CHAPTER 4

Household-level data in a CFSVA

Much of the information used in the CFSVA comes from data collected at household level for the purpose of the CFSVA. Usually secondary data is not sufficient because some key indicators are missing from it is. It is rare to find recent information on household dietary diversity, coping strategies, income and expenditures, and nutrition status all in the same secondary data set, hence primary data are collected.

4.1 SAMPLING IN A CFSVA

Sampling is the methodology by which specific individuals, households, and communities are selected to be surveyed as part of the CFSVA. Sampling is a highly technical activity, and it is critical that the sample design be carefully undertaken. The most common mistake found in CFSVAs and many other data collection exercises is to make errors when designing the sample. Most field offices will require special support from Headquarters or a specialized consultant with sampling expertise. This section focuses on approaches typically used in the primary data collection of a CFSVA.

Why draw a sample? The alternative would be to obtain information on all households, as in a population census. This would provide a very accurate "snapshot" of the population at a particular moment in time. Even groups that were numerically small (and hence possibly missed in a survey) would be counted. However, a census is usually much more expensive than a survey, and processing and cleaning census data would be enormously time-consuming.

When we draw a **random sample** of that same population, we can infer the findings of this sample to the entire population with a known degree of precision. Hence, a smaller sample allows the researcher to devote extra effort to ensure the information obtained is accurate, and it allows for more detail: a CFSVA requires an intensity of interview or observation that cannot be carried out in a census.

Therefore, issues of cost, time, precision, and quantity of data all suggest that a survey is preferred to a census. CFSVAs typically require the conducting of surveys, which have several steps:

- Decide on the sample unit (n), for example, a village or household;
- Determine the "universe" (e.g., rural part of country A);
- Construct a sampling frame (list of all villages in country A; list of all households in a village);
- Decide on the sample size (N); and
- Choose the sample (sampling).

4.1.1 Key terms and concepts[18]

4.1.1.1 Sampling

The term sampling refers to the selection of a limited number of individual units of analysis (denoted as n) from a population of interest (denoted as N) with the purpose

18. WFP, ODAV (VAM), December 2004. *Thematic Guidelines: Sampling Guidelines for Vulnerability Analysis.*

of inferring something about that population from the individual units selected in the sample. Almost all CFSVAs have a primary data collection component. Households and individuals (e.g. mothers and children) are the most common units of analysis in a CFSVA survey.[19]

There are two broad categories of sampling relevant to CFSVAs: probability sampling (also called formal sampling) and non-probability sampling (also called informal sampling).

4.1.1.2 Probability sampling

Probability sampling relies on probability theory to draw statistical inferences about the population of interest from a randomly selected sample. Because probability sampling employs random selection techniques, it is more objective than non-probability sampling. Probability sampling also allows, besides making inferences for the population, for the degree of error around the food security estimates to be quantified.

Probability methods are appropriate when the objective of the assessment is to determine the percentage or number of people who are food insecure.

Example From an exhaustive list of all households in the peri-urban area of Port au Prince, Haiti, 200 households were randomly selected, by assigning numbers to each household and then randomly drawing 200 numbers. An assessment employing probability sampling methods estimates that 28 percent (95 percent confidence interval of +/- 6 percentage points) of households in the peri-urban areas outside Port au Prince consume fewer than two meals per day. In other words, based on a sample survey, we are 95 percent sure that the estimated percentage of households in the peri-urban area outside Port au Prince consuming fewer than two meals per day is between 22 percent and 34 percent.[20]

The types of probability sampling discussed in this chapter include:

- simple random sampling
- systematic sampling
- stratified sampling
- two-stage cluster sampling
- multi-stage sampling

Probability sampling is strongly recommended for all CFSVAs, even in purely qualitative studies.

4.1.1.3 Non-probability sampling

Non-probability sampling relies on a more subjective means of inferring something about the population of interest from a sample. It is not based on statistical theory, and without a statistical basis it is impossible to assess precision and reliability (accuracy)

19. By contrast, nutritional surveys that collect anthropometric data normally treat individuals within households as the unit of analysis. Combined food security and nutritional surveys may use a combination of household- and individual-level analyses.
20. This range estimate is known as a *confidence interval* and is discussed in detail in section 4.1.3.

of estimates. Sample households or individuals are selected because there is reason to believe that they "represent" the population well or that they are well positioned to provide information about the population (as with key informants). The inherent subjectivity and bias associated with non-probability methods are both its strengths and its weaknesses.

Example: To understand the flow of livestock from southern Somalia into Kenya, in-depth discussions are held with a few strategically selected traders (purposive, non-probability sampling). In this case, it makes more sense to select individuals who are knowledgeable than to randomly select individuals who may not know how cross-border trade networks work.

Non-probability sampling methods are appropriate for meeting many of WFP's information needs. Specifically, they are widely used for selecting communities/villages for qualitative studies.

Non-probability sampling is rarely used in CFSVAs for household data collection, and therefore is not covered in this chapter. However, it is frequently used in qualitative data collection, such as focus group and community discussions. Refer to Chapter 5, on community data collection for guidance on sampling in these circumstances.

4.1.1.4 Sampling frames

A sampling frame is an exhaustive list of all sampling units[21] and their physical locations within the population of interest (N) from which the units that will be sampled are selected. The purpose of constructing a sampling frame is to ensure that each household within the population of interest has an equal or known probability of being randomly selected for inclusion in the sample. Random selection of sampling units from a sampling frame allows for estimates from the sample population (n) to be generalized to the larger population of interest (N) defined by the sampling frame.

Sampling frames ensure that every household in the population of interest has an equal chance of being included in the sample.

In practice, sampling frames that are 100 percent complete and accurate do not exist. However, the sampling frames constructed for CFSVAs should be as accurate and complete as possible, but should rely primarily on existing data sources rather than primary data collection.[22] Government census data or demographic data from other surveys are among the most useful sources for constructing sampling frames.

It is important to be transparent about groups or areas that will be intentionally left out of the sampling frame because population (N) level estimates generated by the

21. See section 4.1.1.6 for a detailed explanation.
22. In this instance, primary data collection refers to population data collected in the field by WFP for the purpose of constructing a sampling frame. By contrast, secondary data refers to existing data collected for another purpose that can be used to construct a sampling frame.

sample population (n) do not apply to these groups. Security is perhaps the most common reason for intentionally excluding groups or areas. Some studies may exclude urban areas purposely. However, some individual households or villages are often omitted from the sampling frame unintentionally. Although estimates derived from the sample population (n) cannot be used to generalize about these households, a limited number of chance omissions will not undermine the validity of findings.

4.1.1.5 Sources of information on sampling frames and research design

Secondary data may also be helpful in developing a sampling frame for primary data collection. Four common sources of data of this type are the National Population and Housing Census, the Demographic and Health Survey, the Multiple-Indicator Cluster Survey (MICS), and the Living Standards Measurement Survey (LSMS). These can inform questionnaire design and validate the results of the CFSVA. In cases where data collection for such surveys coincides with the CFSVA, it may be useful to coordinate activities so as to avoid duplication of effort, provided that the sampling and coverage issues can be satisfactorily resolved.

National Population and Housing Census information should be used to develop the sampling framework, since the demographic information often includes rural/urban classification, gender, age, disability, shelter, education level, and migration status by the smallest administrative unit. However, care should be taken to ensure that the census results are still valid (no more than 10 years old) and that no extraordinary events have occurred (wars, conflicts, environmental disasters) that could have significantly changed the census findings.

A review of the latest household consumption and expenditure surveys and of the agricultural census is also very important. These data and analyses assist in defining the critical regional baselines, including long-term average production and consumption patterns. The income, price, supply, and demand elasticity generated from these surveys are extremely useful in determining the expected income, price, and substitution effect for food commodities.

The IMF Generalized Data Dissemination System (GDDS) is also a good source of information about the type of economic, financial, and socio-demographic data available in the country, along with its characteristics, quality, access, and integrity.

4.1.1.6 Primary and ultimate sampling units

The sampling units listed in the sampling frame are the primary sampling units. In some rare cases, such as long-term refugee camps or countries in which a detailed census has recently been conducted, a reasonably accurate sampling frame of all households and their locations is available or can easily be constructed. In these cases, households listed in the sampling frame are both the primary sampling units and the desired units of analysis (also known as ultimate sampling units).

Households are the most common ultimate sampling unit in food security assessments. Villages are the most common primary sampling unit.

However, in most cases a complete list of households for a population of interest is unavailable and would be costly and time consuming to construct. Even if a complete list for the population of interest were available, the cost of visiting households dispersed all over the region of interest would be excessive. In these cases, the sampling frame is constructed at the lowest aggregation of households for which accurate information on the existence, location, and relative size[23] of aggregates is available. In rural settings, this aggregation is often villages, such that an exhaustive list of villages (primary sampling units) within the population of interest can be constructed. In urban settings, neighbourhoods or blocks often provide a suitable aggregation of households and can be used when constructing a sampling frame. Households (the most common unit of analysis in CFSVA surveys) remain the ultimate sampling units.[24] Several options exist for choosing households for inclusion in the sample when the primary sampling units are an aggregation of households such as a village or neighbourhood/block. Choice of a method of household selection is driven by the information available and time/cost constraints. Guidance on choosing an appropriate household selection method is described in detail under each of the sampling methods described in section 4.1.2.

4.1.2 Choosing an appropriate sampling method

A variety of probability and non-probability sampling methods exist to suit different situations encountered in the field. The most commonly used sampling methods for CFSVAs are one or more of the following: simple random sampling, systematic sampling, cluster (or area) sampling, two-stage cluster sampling, and/or stratification. The household survey of a CFSVA typically uses a stratified two-stage cluster sample.

4.1.2.1 Simple random sampling

As the name implies, simple random sampling (SRS) is the most straightforward of the probability sampling methods. A simple random sample involves the random selection of households from a complete list of all households[25] within the entire population of interest (e.g. sampling frame). Households are therefore both the primary and ultimate sampling units. Simple random sampling has a statistical advantage over other sampling methods[26] and requires a smaller sample size (approximately half the size required for cluster or two-stage cluster sampling). However, an exhaustive population list is required, and the cost of visiting geographically dispersed households may be high.

When to apply simple random sampling

In practice, household-level sampling frames are rarely available. However, assessments conducted in long-term refugee camps or areas in which a census has

23. The utility of size estimates is discussed in detail under sections 4.1.2.4 and 4.1.2.5 Cluster Sampling, Two-Stage Sampling, and Multi-Stage Sampling.
24. In rare cases it may be necessary to have multiple levels of sampling units. For example, if no information on villages and their location is available, a higher aggregate, such as a district, may be used. In this example, district is the primary sampling unit (PSU), villages are the secondary sampling unit (SSU), and households (the desired unit of analysis) remain the ultimate sampling unit (USU). A more detailed discussion of this issue is provided in section 4.1.2.4.
25. It is rare to find a complete list of all households to construct a sampling frame for simple random sampling.
26. Systematic sampling shares this advantage.

recently been conducted may provide enough information at the household level to construct one.

For CFSVAs, simple random sampling is almost never applied (except sometimes within a cluster; see section 4.1.2.3 on cluster sampling and selecting households within a cluster).

Despite the statistical advantage and reduced sample size requirements, the existence of a household-level sampling frame does not mean that simple random sampling is always the most appropriate method. Because households are selected randomly from the population, the list of households included in the sample can be widely dispersed and may require visiting a large number of villages to collect the sample.

By comparison, cluster and two-stage cluster sampling limit the number of villages to be visited and may present a logistical advantage over simple random sampling. When the area being covered by an assessment is large, cluster or two-stage cluster sampling may be more cost effective despite the larger sample size requirements.

How to apply simple random sampling

Step 1. Each household in the sampling frame is assigned a unique number between 1 and the total number of households in the sampling frame.

Step 2. A randomization method is then used to select households for inclusion in the sample.[27] Microsoft Excel can also be used to generate random numbers, and even the serial numbers on currency can be used.

Step 3. Next, selected households are mapped to facilitate data collection. The data collection team must also have a household replacement strategy for when (a) a household cannot be located (due to inaccurate information in the sampling frame) or (b) an appropriate respondent is not available.

Step 4. Replacement households can be preselected prior to data collection by identifying the next household in the sampling frame. Alternatively, a protocol[28] for replacing households in the field can be agreed upon prior to data collection. Examples include choosing the next closest household or spinning a pencil in front of the absentee household to select a transect line and choosing the first house encountered in that line as the replacement household. The means of household replacement is less important than the uniform application of the procedure chosen.

Example of applications of simple random sampling

A food security assessment in a **Western Tanzania** refugee camp housing Congolese refugees requires a sample size of 400 households. A list of all households within the camp is available from UNHCR, along with maps locating each household within a block and each block within the camp.

(cont...)

27. The total number of households to be randomly selected from the sampling frame is determined by the sample size requirements (see section 4.1.3).
28. The protocol should be written and provided to each enumerator for reference during data collection.

(...cont)

Each household is assigned a number between 1 and 5,050 (the total number of households in the camp). A random numbers generator (www.randomizer.org, or the RAND function of Excel can be used) is used to select four hundred households. The selected households are then mapped. The workload is divided among four data collection teams, with each team given a mapped area containing approximately 100 households.
Given the proximity of households to one another within the camp, data collection teams are able to walk between selected households. Households that are non-existent or that do not have a suitable respondent available at the time of data collection are replaced by the closest household to the mapped location of the original household.

4.1.2.2 Systematic sampling[29]

Systematic sampling shares the same information requirements as simple random sampling. In contrast to random selection, this method involves the systematic selection of households from a complete list of all households within the population of interest (e.g. the sampling frame). Once again, households are both the primary and ultimate sampling units. Like simple random sampling, systematic sampling has a statistical advantage over other sampling methods and requires a smaller sample size than cluster sampling (approximately half the sample size required for cluster or two-stage cluster sampling).

When to apply systematic sampling

In practice, household-level sampling frames are rarely available. However, assessments conducted in long-term refugee camps or areas in which a census has recently been conducted may provide enough information at the household level to construct one.

For CFSVAs, systematic sampling is almost never applied (except sometimes within a cluster; see section 4.1.2.3, on cluster sampling and selecting households within a cluster). Care must be taken to assess what patterns, if any, exist in the sampling frame. If the ordered pattern has any relation at all to food security, simple random sampling must be applied.

4.1.2.3 Two-stage cluster sampling

In practice, (stratified) two-stage cluster sampling is used in almost all CFSVAs. The combination of minimal information requirements and logistical ease make this method particularly well suited to many of the scenarios encountered during CFSVA surveys.

As the name implies, two-stage cluster sampling is a variant of cluster sampling. A cluster is simply an aggregation of households that can be clearly and unambiguously defined (Magnani 1997). For CFSVA surveys in rural areas, villages are the most common cluster used in sampling. For urban studies, blocks or neighbourhoods may be more appropriate. Two-stage cluster sampling involves selection of a limited

29. Scientific researchers will insist that the only correct way is through random sampling (instead of systematic sampling) because there is always a possibility of "hidden patterns" in the list of households, which may lead to a bias when applying systematic sampling.

number of villages (usually between 25 and 30 in CFSVAs) in each stratum (non-stratified samples have only one stratum). Two-stage cluster sampling uses a second step to select a limited and fixed number of households within each selected cluster. The number of households per cluster varies, but is usually between 8 and 20 for CFSVAs. A 30-by-30 cluster sample is a common form of two-stage cluster sampling, often used in nutrition surveys, where 30 households are selected in each of 30 villages.

Cluster sampling in the CFSVA always uses a "probability proportional to size" selection of clusters. This means that a village with 500 households is 5 times more likely to be selected than a village of 100 households. This ensures that all households, whether from a small or a big village, always have an equal probability of being selected.

When to apply two-stage cluster sampling

The information needed to construct a list of all households in the population of interest (e.g. household-level sampling frame) is often unavailable, and such a list would be time-consuming and expensive to construct. Therefore, a sampling frame is constructed at the lowest aggregation of households for which information is available (often villages, neighbourhoods, or blocks).

Even when a household-level sampling frame does exist, using a random or systematic sampling method is likely to produce a geographically dispersed sample (see sections 4.1.2.1 and 4.1.2.2). Therefore, a large number of villages may need to be visited to select a relatively small number of households.

To reduce the financial costs and time needed to conduct an assessment, particularly one covering a large physical area, a decision may be made to use a two-stage cluster sampling. Two-stage cluster sampling reduces costs and time because it limits the number of villages/neighbourhoods/blocks to be visited and the number of households to be interviewed in each selected village/neighbourhood/block. However, the precision of the results obtained may suffer. For most assessments, the sample size required for a two-stage cluster sampling approach is approximately twice that required for a simple random or systematic sample.[30]

Two-stage cluster sampling is used in nearly all CFSVA sampling approaches.

Example It is determined that the minimum sample size[31] for each rural stratum in a rural CFSVA in Ethiopia is around 180 households. Since we use cluster sampling, assuming a design effect of 2 (180 x 2 =), 360 households are required. Although there has not been a recent census, a reasonably accurate list of all villages in each stratum exists. Looking at one particular stratum, there are 150 villages in total, and their approximate size is available through the government's statistics department. Villages range in size from 20 to 300 households, and on average contain 150 households. At the first stage of selection, 30 villages are randomly selected (with probability proportional to size) for inclusion in the assessment. At the second stage of selection, 12 households are selected within each of the 30 villages, a total sample size of n = 360 (e.g. 30 x 12 = 360).

30. This is due to the design effect of using a cluster sampling methodology. This issue is discussed in detail in this chapter. A design effect of around 2 is common for many indicators if the cluster size is around 10 to 15 households.
31. Assuming simple random sampling.

How to apply two-stage cluster sampling

Two-stage cluster sampling requires three distinct steps: (1) defining clusters and constructing the sampling frame; (2) choosing clusters for inclusion in the sample; and (3) choosing households from within selected clusters for inclusion in the sample. As with cluster sampling, each of these steps involves a number of intermediate steps.

Selecting clusters to include in the sample

Step 1a. The first step in (two stage) cluster sampling is defining the aggregation of households that will be used as "clusters." The following criteria are helpful for defining appropriate clusters[32]:

- Aggregations should be pre-existing and recognized. Villages, blocks, neighbourhoods, and census blocks are good examples.
- Aggregations used for clusters should be as unrelated to food security as possible. Unlike stratification – in which households are categorized into sub-groups on the basis of criteria related to food security such as livelihoods and land-use zones (e.g. homogeneity) – the aim of clustering is just the opposite (e.g. heterogeneity). Ideally, each cluster should contain households that reflect the diversity found in the entire population of interest (in terms of food security–related factors such as livelihoods and land use). For the majority of CFSVA surveys, the use of administrative aggregations (such as villages) as clusters will most closely approximate this ideal.
- Information on the size of the cluster (number of households or population size) is available.

Where population estimates are unavailable, key informants can be used to provide rough/relative estimates for all villages in the sampling frame (e.g. very large, large, medium, small, very small).

Step 1b. Next, assemble the sampling frame. For stratified samples, a separate sampling frame must be developed for each stratum (e.g. sub-groups defined by stratification criteria). Microsoft Excel or similar spreadsheet software is useful, though a simple table can also be used. In the first column, list each cluster. In the second column, list the size of the cluster (either population or number of households). If you are using rough estimates from key informants, use relative size codes. The two-column table under Step 1b provides example codes.

Cluster Size	Code
Very large	5
Large	4
Medium	3
Small	2
Very small	1

Step 1c. Use the third column to list the cumulative size values for all clusters. The cumulative size value for cluster B is the sum of clusters A and B. The cumulative size value for cluster C is the sum of clusters A, B, and C... and so on.

32. The first, third, and fourth criterion were adapted from the FANTA Sampling Guide (Magnani 1997).

Example Sampling Frame with Cluster Population Estimates

Cluster	Size (pop.)	Cumulative Size
A	50	50
B	125	175
C	35	210
D	20	230
E	80	310
F	20	330
G	25	355
H	40	395
I	25	420

Example Sampling Frame with Key Informant-Generated Cluster Size Estimates

Cluster	Size (category)	Cumulative Size
A	3	3
B	1	4
C	5	9
D	2	11
E	1	12
F	1	13
G	4	17
H	5	22
I	3	25

Step 2a. The next step is to decide how many clusters will be included in the sample. As indicated above, 25 to 30 clusters per stratum are typical for most settings (non-stratified samples have only one stratum). The recommended size of the clusters is between 8 and 20 households. The recommendation of 30 clusters per stratum is somewhat arbitrary, but provides a commonly used and technically sound standard. Choosing the most appropriate number of clusters requires striking a balance between technical and logistical considerations. A bare minimum of 20 clusters (preferably 25) per stratum provides a lower limit for surveys where cost and time considerations are major constraints.[33] Typically, CFSVAs have around 25 or 30 clusters per stratum, and if increasing the number of clusters (and decreasing the sample size per cluster) does not affect the survey logistics or cost, this is a preferable option, as it decreases the design effect with a constant sample size.

33. Reducing the number of clusters to below 20 requires a technical assessment of the expected inter-cluster heterogeneity and intra-cluster homogeneity and should not be done without appropriate technical guidance. Fewer than 20 clusters may be possible in samples in which stratification produces a large number of sub-groups (e.g. strata are very homogenous on factors related to food security, reducing the range of heterogeneity within and between clusters within particular strata).

Example A CFSVA survey in rural India required a sample size of 300 households in each of 5 strata (sub-groups defined by land-use zones) for a total sample size of n = 1,500. Information from the government allows for the use of villages as clusters. The following options were considered for each of the 5 strata:

30 clusters of 10 households each (n = 300)
25 clusters of 12 households each (n = 300)
20 clusters of 15 households each (n = 300)

Since there are 5 strata, a decision is made to take the minimum acceptable number of clusters to reduce the number of vehicles and other costs associated with the assessment. The total number of clusters/villages to be visited is 100 (25 clusters in each of 5 strata), for a total sample size of n = 1,500 (12 households in each cluster). This worked well because 1 team of enumerators, with 1 vehicle, was able to interview 12 households in a village in one day, with enough time left to travel to the next location. Fifteen households was too many to interview in a day, and 10 households would have left extra time but not enough to start on a different village in the same day. Additionally, the limited impact on the survey's design effect, by increasing the cluster size from 10 to 12, and the number of clusters from 30 to 25, was considered acceptable in this case.

Step 2b. Use the number of clusters, number of households per cluster, and number of days allotted for data collection to determine the number of enumerators/data collection teams required. Since adding a few more households per village is logistically easier than having more villages of smaller size, constraints on the number of enumerators and teams available may suggest using the compromised (25) or minimum (20) number of clusters. However, a serious attempt should be made to find additional enumerators or add data collection days before reducing the number of clusters. A pre-test or experience in other surveys will help to estimate the number of interviews a data collection team of reasonable size (3 to 5 enumerators) can complete in a day.

Example Continuing from the Indian example (with 25 clusters in each of 5 strata, with 12 households taken per cluster for a total sample size of n = 1500), it is estimated that each enumerator can complete 4 interviews per day. Therefore a team of 3 enumerators (with each doing 4 households) and 1 team leader (working on the community questionnaire and supervising household data collection) can complete 1 cluster per day. Twenty days have been allotted for data collection. Since there are 125 clusters (25 x 5), 8 such teams could complete the work in about 16 days, with 4 extra travel days (or more, depending on the distance between clusters), it is estimated that 8 teams will be needed (24 enumerators plus 8 team leaders).

Step 2c. Clusters are then randomly selected from the cluster-level sampling frame. Cluster population figures are used to select clusters with a probability proportional to size (PPS), meaning that larger clusters have a higher probability of selection. As indicated earlier, key informants can be used to provide rough estimates where existing information on cluster size is unavailable.

Box 4.1: Probability proportional to size (PPS)

The purpose behind selecting clusters "PPS" is to ensure that each household in the population of interest, whether from a large or small village, has an approximately equal probability of selection. To approximately equate probability of household selection at the second stage, large villages must have a higher probability of selection at the first stage. Selecting clusters without PPS leads to households having different probabilities of selection; households from small villages are overrepresented. Such samples are non-self-weighting, and can complicate analysis (Magnani 1997).

Example The required sample size for a survey in rural northern Uganda is 250 households. Information on the location and approximate size of villages is available through the government. A total of 75 villages is listed in the cluster-level sampling frame. Twenty-five villages will be chosen for the sample and ten households will be taken in each of the selected villages for a total sample size of n = 250.

Random selection – Generate 25 random numbers between 1 and the total cumulative population (or household or size code values). The clusters containing each of the cumulative numbers selected are included in the sample. Statistically, if a cluster is selected twice in this example, 20 households should be taken in that cluster. However, in practice in CFSVAs, this is not always applied. To avoid this problem, one should be cautious when creating clusters. If clusters sizes are often large, they should be subdivided so that they are not double-selected.

Systematic Selection – To determine the sampling interval (SI), divide the total cumulative size indicated in the last cluster listed in the sampling frame by the number of clusters to be selected (25). Generate one random starting number between 1 and the sampling interval. The cluster containing the cumulative number selected is the random starting household.

Example
111 is the randomly selected "first household" selected from the range 1–200.04 (range defined by the sampling interval). This cumulative size corresponds with cluster B in the example here:

Cluster	Size (pop.)	Cumulative Size
A	50	50
B	125	175
C	35	210
D	20	230
E	80	310
F	20	330
G	25	355
H	40	395
I	25	420
J	100	520
etc.	etc.	etc.

To select the second cluster, add the sampling interval to the cumulative size given by the random start. The cluster containing the product is the second cluster. To select the third cluster, add the sampling interval to the cumulative size used to select the second cluster... and so on, until 25 clusters are selected.

Example Second household (200.04 + 111 = 311.04) located in cluster F. Third household (200.04 + 311.04 = 511.08) located in cluster J, and so on.

A common trick when selecting clusters systematically is to order the villages (or other cluster unit) by their location within the strata. For example, if in a survey where the main stratification is by province, and a two-stage cluster sample is being drawn in each province, the list of villages can be ordered by geographic area, such as district and livelihood zone, before taking the systematic PPS sample of villages. This can reduce the chance (even if small) of having all villages located within one district (or livelihood zone, or other geographic region) in a province, even if that province has three districts.

Selecting households within selected clusters[34]

Several options exist for selecting households within selected clusters. Each option can be applied regardless of whether the clusters were selected randomly or systematically (Step 2c). Two options are listed here in order of preference; however, the second option is cheaper and faster than the first. Choosing the right method for household selection will vary by assessment. Assessments should strive to use the preferred method (Option 1), choosing Option 2 or an alternative method only when absolutely required due to logistical, time, and resource constraints.

Option 1. The ideal household selection method involves constructing a sampling frame of all households within the selected clusters. Where clusters are small in size, this approach is manageable. However, it will be costly and time prohibitive when the clusters are large. Once the sampling frame has been constructed, follow the guidance given for simple random sampling or systematic sampling for selecting households for inclusion.

Example An assessment is being carried out in rural Bangladesh. Villages will serve as clusters. Thirty villages have been selected for inclusion in the sample in each of two strata for a total of 60 villages. Ten households will be selected in each village for a per-stratum sample size of n = 300 and a total sample size of n = 600. Upon arrival in each selected village, the data collection team maps the village, giving each household a unique number (no two households can have the same number). In the first cluster there are 35 households, such that the households are numbered 1 to 35.

Option 1a. One option is to select households randomly. Write down each household number (1 to 35) on a slip of paper and put them in a hat. Shake the hat and then select 10 slips of paper. The number on the slip of paper corresponds with the household to be interviewed.

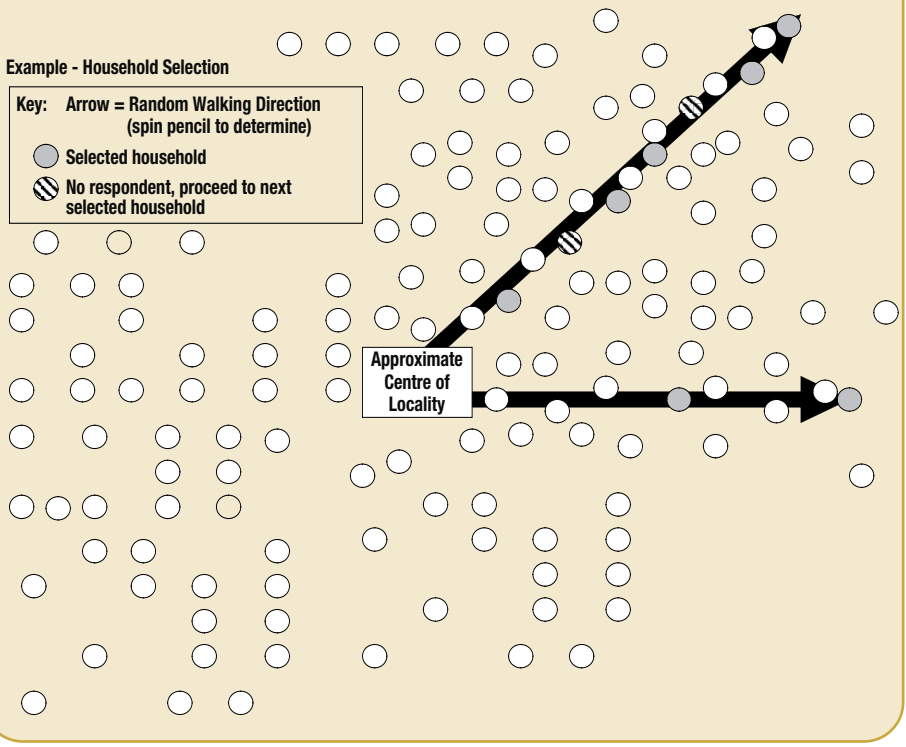
34. This section borrows heavily from the procedures outlined in the FANTA Sampling Guide (Magnani 1997).

Letting members of the community choose from the hat provides an excellent means of involving them in the process, helps them to understand the meaning of "random selection," and avoids scenarios in which village leaders attempt to dictate which households are to be interviewed.

Option 1b. A second option is to select households systematically. A sampling interval of 3.5 is calculated (35 HHs in the village divided by 10 HHs needed for the sample) in this example. Household 2 is selected as the random starting household (chosen in the range of 1 to 3, since 3.5 contains a decimal). The sampling interval of 3.5 is added to the random start to select the second household (5.5, rounded up to household 6). Add the sampling interval again to get the third household (5.5 + 3.5 = 9), and so on.

Option 2. The second option for selecting households is the most rapid, but also the less preferred method. This method is sometimes used in the Expanded Programme on Immunization (EPI) surveys and in UNICEF anthropometric surveys. Once the data collection team arrives in the cluster, the approximate middle of the cluster is identified. A pencil or bottle is spun to select a random walking direction (also called a transect line). The data collection team then counts the number of households encountered along the transect line between the centre and the perimeter of the cluster. This number is then divided to determine the interval at which households along the transect line will be selected.

Figure 4.1: Household Selection



When the transect line contains fewer than the number of households required, all households in the line are included in the sample and the data collection team returns to the centre of the cluster to pick a second random walking direction, and the process is repeated. If a household without an appropriate respondent is encountered, skip it and proceed to the next selected household. This may require returning to the centre and repeating the process for transects with fewer than the number of required households. This method usually results in a bias, because households from the centre of the village can be overrepresented. Additionally, enumerator teams tend to bias themselves toward transects along main roads or paths.

Example An assessment was carried out in Tambura District in Southern Sudan. Villages served as clusters. Thirty villages were selected in each of 2 livelihood zones, with each representing a stratum. Ten households were selected in each village for a per-stratum sample size of n = 300 and an overall sample size of n = 600. Upon arrival in each selected village, the data collection team asked two key informants to help locate the centre of the village. A pencil was spun to pick a random walking direction (transect). The number of households encountered when walking from the centre of the village to the perimeter was 20. Therefore, every other household was selected for inclusion in the sample.

In two households, an appropriate respondent was unavailable. Therefore, the data collection team was required to repeat the process by returning to the centre, picking a transect line, dividing the number of households in that line by 2 (the number of replacement households needed). With 8 households in that transect, this resulted in every fourth household in the second transect line being sampled.

4.1.2.4 Multi-stage sampling

In the majority of CFSVAs, a two-stage cluster sampling methodology is used. However, on rare occasions a multi-stage method may be required.

Multi-stage sampling is an extension of the two-stage random sampling (e.g. three or more stages). For example, accurate information may exist only at the division level, necessitating three (or more) sampling stages:

Stage 1. Random selection of villages
Stage 2. Random selection of households within selected villages
Stage 3. Random or systematic selection of household members within selected households

The design effect, and therefore sample size requirements, increase considerably with each additional sampling stage. Therefore, multi-stage sampling (where districts are sampled, and then, within them, villages, and then, within those, households) is not recommended.

A common mistake when designing a survey is, for logistical reasons, sampling a limited number of districts (one or two) for each province first (stage 1), and in these districts sampling a number of villages (stage 2), and from those villages, sampling 10 households for interview. Such a design will have a huge design effect, and hence very imprecise

population estimates. If only one district is sampled from a province, no generalized statements about that province can be made.

4.1.2.5 Stratification or stratified sampling

Stratification, or stratified sampling, involves dividing the population of interest into sub-groups (e.g. strata) that share something in common based on criteria related to the assessment objectives.[35] Stratification is used when separate food security estimates are desired at a predefined, minimum level of precision for each of these sub-groups. When used appropriately, stratification can increase the precision of overall food security estimates for the population of interest.

Stratification by administrative boundaries allows for separate estimates to be generated for disaggregated areas within a population. For example, a national sample may be stratified by district in order to ensure the precision of food insecurity estimates at the district level for comparative purposes.

However, stratification is most effective when it is used to define sub-groups within the population that share characteristics related to vulnerability or food security. Livelihoods and land-use zones are examples. If there are distinct livelihood zones in the area where the CFSVA is to be conducted (e.g. agricultural, pastoral, agro-pastoral groups), they can be used to stratify the population. Defining groups in this way serves two functions. First, administrative boundaries rarely correspond with household characteristics related to food insecurity and estimates for administrative aggregations are likely to mask meaningful differences between sub-groups. Second, defining sub-groups for stratification using criteria related to vulnerability or food insecurity improves the precision of both sub-group and overall food security estimates.[36]

Stratified sampling is a key component of all CFSVA sample designs, and is used for comparing sub-groups within the population of interest, an important objective of any CFSVA.

Example The estimated percentage of food-insecure households for Garissa, Kenya, a rural district containing both an area with primarily nomadic pastoralists and one with primarily sedentary farmers (livelihood zones), is 35 percent (+/- 5 percentage points). However, this average at the district level masks the fact that 70 percent of pastoralists are food insecure, while only 10 percent of sedentary farmers are food insecure.
Stratified sampling requires that each sub-group (stratum) be mutually exclusive, meaning that every household in the population of interest must be assigned to only one sub-group. The strata should also be collectively exhaustive, meaning that every household in the population of interest must belong to a sub-group.

35. The purpose of stratification is to define homogenous sub-groups within a heterogeneous population for comparison and, to a lesser extent in CFSVAs, to increase the overall precision of estimates derived from the sample.
36. Stratification by sub-groups defined by criteria related to food security results in more homogenous groupings in terms of food security outcomes. The result is an increase in the precision/accuracy of estimates for each sub-group and of the combined overall estimate for the population due to reduced sampling error. By contrast, stratification by administrative boundary is likely to result in heterogeneous groupings similar to the heterogeneity found in the overall population under study.

In many CFSVAs, two separate geographic stratification systems are used simultaneously. For example, both administrative boundaries and livelihood zones could be used to define the strata. It is important to match what is commonly used in the country, to allow for comparability.

Since information related to food security and vulnerability is most often found for administrative aggregations (districts, divisions, provinces, departments, etc.) or agro-ecological zones, stratification in a CFSVA is always done on a geographic basis. We may prefer to stratify by population group (livelihood groups, gender, wealth groups).[37] However, lack of data almost always makes this impossible.

If it is the intention to report on every cross-section of the two stratification systems, which entails the inclusion of additional villages and households in the sample, each additional sub-group (e.g., stratum) represents an increase in cost and time required to conduct the assessment. If the reporting domain is each stratification system separately (and not the cross-sections), cost increase is limited. Therefore, cost and time constraints will figure heavily into if and how a sample can be stratified. If, for example, the sample size required for a province level of estimate at a reasonable level of precision is 200 households, stratifying the province into two livelihood zone sub-groups would require applying the same sample size to each of the two livelihood zones if the same level of precision were desired for each sub-group (200 x 2 = 400).

Example A food security assessment in Country ABC was originally designed to yield district-level estimates for four districts (four strata). The estimated sample size required was 400 households per district for a total of 1,600 households.
Upon further reflection, the country office wanted results reported by major land-use zones within each district (requiring stratifying by two criteria). Land-use maps suggested that two of the districts had four land-use zones and the other two districts had three land-use zones, for a total of 14 land-use zones/district strata. To keep the same precision in each of these 14 combinations, a sample size of 400 is needed in each zone, increasing the sample size to 5,600 households.
However, another option was to increase the sample with just 450 HHs to 2,050 HH. As a result, each of the four land-use zones also had a sample size of at least 400 HH. Data could now be reported either by district or by land-use zone.

Given these practical limitations, it will not be possible to stratify a sample by every comparison you wish to make during analysis, particularly household-specific characteristics (rather than geographic areas). But if a sub-group is well represented in the population, it is likely that a sufficient number of households within that sub-group will be randomly selected. As a result, a fairly precise estimate of the food security status of the sub-group can be generated during analysis without pre-stratifying the sample.

All CFSVAs use some sort of stratified sampling, usually based on geographic areas (administrative boundaries and/or livelihood or food security zones). A best practice is

37. Of course, this does not prevent reporting according to these individual household characteristics, such as livelihood.

to design the sample so that it equally satisfies both an administrative stratification and a food security zone stratification. This means that the analysis can provide aggregates for both stratifications. In the report, one can highlight the stratification that best explains the observed differences in a particular indicator.

4.1.2.6 Implications of a complex sample design

The sampling strategy must be taken into account in data analysis. The software SPSS assumes a simple random sample, which assumes that each subject (household) has the same probability of being selected, and that the selection of each subject is independent of the selection of any other subject. This is rarely the case with CFSVAs, which typically use a complex sample design using stratification and multi-stage (usually two-) cluster sampling.

4.1.2.7 Error of estimates/design effect

Cluster sampling produces a less precise estimate than a simple random sample. This is referred to as the design effect, which is the number by which the sample size is multiplied to get the same margin of error as a simple random sample. The design effect will be different for each variable, and can be calculated only *post hoc*. In calculating sample size, an educated guess is made at the design effect, often assuming a design effect of 2. However, this is just an assumption; the design effect can be smaller or much larger, depending on the cluster size and the circumstances in the country.

The reason for the design effect is that households in the same village are often similar to each other (i.e. there is an intra-cluster correlation). Twenty households from 2 villages will not reveal as much about the entire population as 20 households from different villages. The higher the intra-cluster correlation, and the more households coming from the same cluster, the higher the design effect.

To calculate all measurements of errors (including confidence intervals, standard deviation, standard error, and variance), SPSS assumes a simple random sample. Only by using the complex samples module of SPSS (which is not included in the standard version) can complex sampling designs be taken into account.

4.1.2.8 Weights

The most important effect of a complex sampling design is the need for weighting. If each household in the sampling frame (and therefore, the resulting sample) has an equal probability of being selected, then no weighting system is needed. If this is not the case, then a weighting system needs to be used. Weights are needed to compensate for the unequal probabilities of a household being included in the sample.

As the goal of a CFSVA is to make estimates to the larger population, the use of weighting needs to be taken into account at each level of a survey, from design to reporting. Statements such as “40 percent of the sample responded that . . .” are a subtle way to get around data that is not representative of the greater population, but also reveal little or nothing about the greater population. For example, the research team is not interested in the food security situation of the 2,000 sampled households; rather, they want to infer conclusions to larger groups and regions and therefore need to make statements such as “40 percent of all rural households are . . .”

For example, in country A, estimates are desired for each of the three provinces, X, Y, and Z. An equal number of households are selected from these three provinces, and all households within each province have the same probability of being selected. However, the populations within the provinces are unequal. This means that when analyses are run on unweighted data, the estimates will be accurate by province, but when estimated by any other stratifying variable or together for a national average, the results will be biased – that is, the numbers will reflect the sample but not the population from which the sample is taken.

Design weights can be described as "the inverse of the probability that a household could be selected." Another definition for the weight is "the number of HHs represented by one sampled household in a substratum." For example, if in a province with 100,000 households 200 HHs are sampled, the weight of each household equals 500 (every household represents 500 HHs from the population); if in another province with 50,000 HH, again 200 HHs were sampled, the weight for that province would be 250 (every household represents 250 HHs of the sampling universe).

Box 4.2: Calculating design weight

$$Ws = \frac{N_s}{n_s}$$

W_S: Design weight in sampling stratum s
n_S: Sample size of sampling stratum s
N_S: Number of households in sampling stratum s

To correct for this unequal sampling probability using weights, a weighting factor must be added in SPSS. Although the weights as calculated in Box 4.2 are correct, in practice the "normalized weights" are used as correction factors.

Box 4.3: Calculating normalized weight

"Normalized weight" can be arrived at by the following formula:

$$w_s = \frac{N_s / N}{n_s / n} = \frac{N_s / n_s}{N / n} = \frac{N_s}{n_s} \cdot \frac{n}{N}$$

w_s: Normalized weight for sampling stratum s
N_s: Number of households in sampling stratum s
N: Total number of households in the entire sampling universe
n_s: Sample size of sampling stratum s
n: Total sample size of all sampling stratums
N_s/N: Proportion of all households living in sampling stratum s
n_s/n: Proportion of sampled households coming from sampling stratum s
N_s/n_s: The design weight in stratum s
n/N: The sampling fraction of the survey

Table 4.1 illustrates an example of normalized weight calculation. The example illustrates that 15 percent of all households of a country live in province A; however 20 percent of the households are sampled from province A. The normalized weight to be applied to households from province A in SPSS is: 15%/20%= 0.75.

Table 4.1: Example of normalized weights

Province	Total households (N_S)	(Ns/N) (%)	Sampled households (n_S)	(ns/n) (%)	Design weight (N_S/n_S)	**Normalized Weight (ws) [(Ns/N)/(ns/n)]**	Weighted number of households
A	30 000	15	300	20	100.0	**0.75**	225
B	80 000	40	450	30	177.8	**1.33**	600
C	50 000	25	400	27	125.0	**0.94**	375
D	40 000	20	350	23	114.3	**0.86**	300
All	200 000	100	1 500	100	133.3	**1.00**	1 500

Once these weights are calculated, new variables in SPSS must be created, and each case will have its design weight and its normalized weight recorded. Using the example in Table 4.1, all the households in Province A will have a normalized weight of 0.75 and a design weight of 100. All households in Province B will have a normalized weight of 1.33 and design weight of 177.8, and so on.

The weights can then be applied in SPSS by activating them. Under "Data" select weight cases, and then select the variable that contains the appropriate normalized weight. In the complex sample procedure of SPSS, the design weight is used.

Box 4.4: How to double-check weights

The weighted total number of all households should be equal to the original unweighted number of households, if normalized weight is to be used.

The unweighted mean of the normalized weights in the data set should equal 1. Think about normalized weights logically: in areas that are very over-sampled (where the sampling fraction is larger – e.g., a province with a small population where the same number of households were sampled as in other, more populous provinces), the weight should be smaller. In areas that are very under-sampled (where the sampling fraction is smaller), the weight should be larger.

In an efficient sample design, the normalized weights should not be very different from 1 (ranging from 0.75 to 1.5 or so). Very large weights (above 2 or so) and very small weights (below 0.5) can decrease accuracy.

4.1.2.9 Considerations in complex samples

A CFSVA typically uses a two-stage sampling design. First, clusters (villages) are sampled, and during the second stage, households are sampled in each cluster. The clusters must be taken into account during analysis, particularly when calculating tests of significance, standard deviations/variations, and confidence intervals. Compensating for clusters in analysis will not alter point estimates, but will change (widen) confidence intervals and variations. Clusters can be dealt with in a variety of

ways. The design effect is influenced primarily by two things (assuming constant sample size): (1) the number of households in each cluster, and (2) the intra-cluster correlation.[38] Another way to think about it is that, keeping constant sample size, more clusters (and therefore fewer households per cluster) means a smaller design effect.

- With a constant sample size, if there is a large number of clusters in the area of estimation (greater than 50 or so) and the number of households in each cluster is small (less than 10 or 15), then the design effect will be small (typically around 2 for many indicators used in CFSVAs). In this case, analysis can be run without considering the design effect (this should be clearly stated in the report). Additionally, statistical tests are more likely to find significant differences when there are none, and confidence intervals will be larger than calculated. Although not ideal, this is also the most common approach.
- If there are a small number of larger clusters in the area of estimation (not a typical situation), complex sample analysis (a procedure in SPSS) must be used.

A typical CFSVA generally uses many clusters (e.g. 250 clusters of 10 households). Typically small cluster size results in a higher number of clusters, which gives a smaller design effect (although it depends on the sample and the indicator). Other surveys, such as nutrition surveys, use 30 clusters of 30 households, which tend to give a design effect of around 2 for nutritional indicators. Compared to other socio-economic or livelihood indicators, nutritional indicators have lower intra-cluster correlation. WFP's experience from Niger shows that 10 households per cluster gives a design effect of 2.0.

When analysing nutritional indicators (stunting, wasting, underweight, etc.) and other methodologically bound and important indicators, the effect of clusters should always be accounted for in the analysis.

Another positive but little exploited effect of a complex sample is stratification. If a set of strata in the survey explain a lot of the variation in the indicator (the variation between strata is large, but within a stratum, small), then using SPSS's complex samples, this can be taken into account, and a greater level of precision can be achieved in the statistical tests.[39] However, the benefits rarely add much value to the results, so this is rarely used in CFSVA analysis.

4.1.2.10 Typical survey design: sampling, weights, and analysis

CFSVAs usually use a complex sample design, using both stratification and a two-stage cluster approach. Usually, areas of estimation have a minimum of 20 or 25 clusters, where each cluster has a minimum of 10 households. Most of the time, these are not self-weighting samples, and need probability weights in analysis.

Considering this, it is absolutely necessary to use the weights in analysis, and their use is strongly recommended to account for the cluster design.[40] However, if the clusters

38. Intra-cluster correlation is a measurement of the relative homogeneity or heterogeneity within clusters as compared to the homogeneity or heterogeneity between clusters.
39. This means a design effect of 0.9 is possible in this case.
40. To account for cluster design, use the module "complex samples" in SPSS; the alternative: simply assuming a design effect equal to 2 is not especially accurate.

are not accounted for, it should be reported in the methodology that any confidence intervals (CI) reported are likely to be wider, and that the tests of significance may indicate significance when there is in fact none.

4.1.3 Determining the appropriate sample size

The aim of this section is twofold: to provide a basic understanding of the factors to be considered in the calculation of sample size, and, more important, to offer easy-to-use[41] sample size guidance for common scenarios found in CFSVA surveys. Only the most basic sample size calculation guidance is given here, which is even more than what is needed in CFSVAs, since almost always the "rule of thumb" sizes of 200–300 HHs per reporting domain is applied.

The choice of sample size formulas depends on whether the key food security indicator (or indicators) of interest for the assessment is mean or proportionate.[42] A primary objective of most CFSVA surveys is to estimate the percentage of food-insecure households within the population. However, some CFSVA surveys will use other indicators expressed as means.

There are many misunderstandings concerning sample size. Perhaps the most common has to do with population size. Except where a population is exceptionally small and a "finite population adjustment" is required, population size has nothing to do with the size of the sample required.

Ultimately, the choice of sample size is almost always driven by practical limitations on time and resources. However, this does not render the calculation of sample size on the basis of technical factors irrelevant. Sample size calculation provides the ideal sample size required to meet the objectives of the assessment. Knowing this is critical to understanding the consequences of deviating from the ideal due to cost and time constraints and allows for informed choices to be made.

4.1.3.1 Sampling when key indicators are expressed as percentages

The formula for calculating the sample size for assessments with key indicators expressed as percentages is:

$$\mathbf{n = (D)(z^2 * p *(1-p))/d^2}$$

Where:

n = Required minimum sample size
D = Design effect (often assumed to be 2, but varies by type of sampling and by indicator)
z = z-score corresponding to the degree of confidence (1.96 if degree of confidence is 95 percent)
p = Estimated proportion of key indicator expressed as a decimal (e.g. 20 percent = .20)
d = Minimum desired precision or maximum tolerable error expressed in decimal form (e.g. +/- 10 percentage points = .10)

41. Guidance is provided that does not require users to make the calculation themselves.
42. The term *proportion* includes percentages and prevalence.

Taken as a whole, the formula can be intimidating, particularly for those unfamiliar with mathematical notation. However, taken separately, each parameter in the formula is relatively easy to define, and automated sample size calculators are available to perform the necessary computation. In addition, recommended sample sizes (not requiring computation) are provided for common scenarios encountered in CFSVA studies.

D The design effect for simple random sampling is equal to 1 (meaning there is no design effect). The design effect for cluster or two-stage cluster sampling is the factor by which the sample size must be increased in order to produce survey estimates with the same precision as with a simple random sample.[43] A typical value for cluster and two-stage cluster sampling is 2, resulting in a doubling of the sample size requirement. However, it may be possible to reduce this value by increasing the number of clusters, and hence having a lesser number of households in each cluster, or when design effect estimates for the same indicator are available from previous surveys.[44] For a given number of clusters, the gains from adding a few additional households are usually minimal. In a CFSVA, typically 10 to 15 households are selected and interviewed in each cluster (a village), with 17 households, instead of 15, interviewed resulting in a minimal gain in precision.

Z Due to the fact that estimates are based on a sample, rather than total enumeration of the population (as in a census), it is not possible to be 100 percent confident that the estimate derived from a sample is a true reflection of the population. The conventional degree of confidence for almost all social research is 95 percent, meaning that if you were to perform the assessment 100 times, 95 of the 100 assessments would yield range estimates known as confidence intervals (e.g. 20 percent +/- 5 percentage points) containing the true population proportion. By contrast, 5 of the 100 assessments would yield confidence intervals that do not contain the true population proportion due to chance. The z-score corresponding with 95 percent confidence is 1.96, which is the standard used in CFSVAs.

p An estimate (in decimal form) of the primary food security indicator of interest allows the sample size to be reduced. Where no reasonably accurate estimate can be found, a default value of 50 percent should be used. This default offers a safe, albeit more expensive, alternative, as the value of 50 percent will yield the largest required sample size.

Since CFSVAs report a variety of indicators (not just percentage of food insecure), it is generally recommended to use the default of 50 percent, knowing that certain indicators with a higher or lower prevalence than 50 percent will have tighter confidence intervals (i.e. more precision).

d The primary technical choice in determining sample size for a non-stratified sample is defining a minimum level of precision (or maximum tolerable error). Precision refers to the degree of error (or confidence interval) around the estimate due to the fact that the estimate is based on a sample.

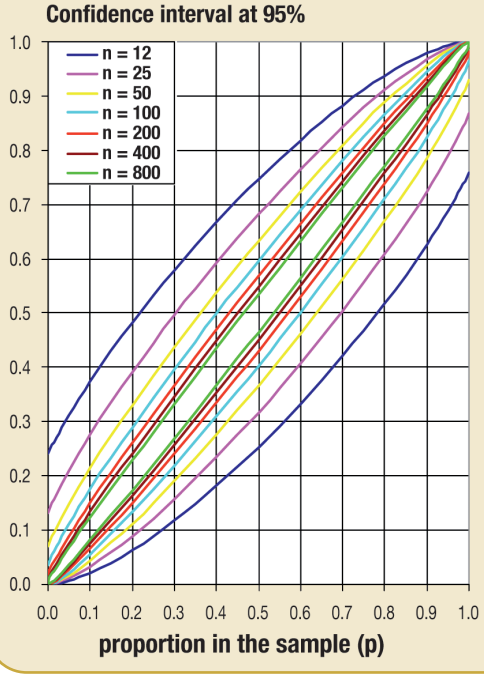
43. See FANTA Sampling Guide for a more in-depth discussion (Magnani 1987).
44. Demographic and Health Surveys (DHS) often have estimates of the design effect of two-stage cluster sampling for food security indicators.

Example It is estimated that 28 percent (+/- 5 percentage points) of households in a rural district in Bolivia consume meat less than once a week. The "+/- 5 percentage points" is the degree of error around the estimate; it defines the confidence interval. The point estimate 28 percent reflects the percentage actually found in the sample population. The range, or confidence interval, of 23–33 percent better reflects the larger population from which the sample was taken.[45] The larger the sample, the narrower the confidence interval.

4.1.3.2 Sample size guidance

Figure 4.2 depicts the intervals that will be obtained with a 95 percent level of confidence, for different simple random sample sizes and different values of an indicator expressed as a proportion. Examples of such indicators are the proportion of food-insecure households, proportion of illiterate household heads, and the global acute malnutrition rate. Figure 4.2 clearly shows that the precision of the estimate is less for proportions around 0.5, and that the higher the sample size, the higher the precision. It also shows that in order to increase the precision appreciably, if there are already 200 units, one has to add a large number of additional units. This graph can be used to gain an idea of what sample size is needed to obtain the desired precision.

Figure 4.2: Confidence intervals for proportion estimates using simple random sampling[46]



Example: The assessment will employ a two-stage cluster sampling method (design effect = 2). The estimated population proportion is 0.4. The graph shows that for a simple random sample of 100 HH, the interval would be 0.31 to 0.50. Hence a cluster design sample with a design effect of 2 would need (100 x 2 =) 200 HH to obtain the same confidence interval. A cluster design sample of 400 households would have a confidence interval of 0.34 to 0.46 by following the lines for n = (400/2 =) 200.

Example: An assessment in West Bank/Gaza will employ a two-stage, cluster sampling method. An estimate for the key food security indicator is 60 percent for the target population. The assessment team decides that the estimate for the population should have a degree of error no larger than 5 percentage points (pp in column) in either direction (+/- 5 pp). This corresponds with the n = 400 line; assuming a design effect of 2, the sample needs (400 x 2 =) 800 HH. (The exact calculation results in a required sample size of n = 762.)

45. As discussed under z, in section 4.1.3.1, the convention for confidence intervals is 95 percent. A comprehensive statement about the estimate given in the example would be "we are 95 percent confident that the true proportion of households in X District, Bolivia, consuming meat less than once a week falls between 23 percent and 33 percent" or "it is estimated that 28 percent (95 percent CI 23–33 percent) of households in X District, Bolivia, consume meat less than once a week."
46. Based on the Wilson score interval. See the formula at: http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval.

Box 4.5: Nutrition surveys vs. food security sampling requirements

Food security analysis has less stringent demands on the precision of a single indicator, since convergence of evidence is used to make conclusions about food insecurity, whereas nutrition indicators (stunting, wasting, and underweight) demand a greater level of precision. A few percentage points difference in prevalence of "food insecurity" is acceptable as a margin of error. However, a difference of a few percentage points in the wasting prevalence can have considerable implications for programmes and responses.

Nutrition survey

Most surveys measure more than one nutrition outcome (wasting, stunting, etc.). In such cases, the sample size for each outcome should be calculated. The largest sample size is used in order to ensure that all outcomes have an adequate sample size.

Many nutrition survey managers select a sampling consisting of 30 clusters, with 30 children or households in each cluster (often called 30-by-30). Such sampling has also been recommended by many organizations in order to ensure sufficient precision when the survey is completed. However, in some situations, the 30-by-30 cluster survey provides more precision than is actually needed. The 30-by-30 sample size assumes that:

- the prevalence is 50 percent;
- the desired precision is -+ 5 percentage points;
- the design effect is 2; and
- 15 percent of the households or children will refuse or be unavailable.

However, in most emergency-effected populations, the prevalence of acute malnutrition is much lower than 50 percent, far less than 15 percent of households refuse or are unavailable, and the design effect may be less than 2. Therefore, for most nutrition surveys where wasting is the primary indicator of interest, the desired precision can be obtained with a sample size substantially less than the 900 included in the 30-by-30 survey. The solution should be to calculate the sample size that provides the desired statistical precision, but not more.[47]

Household food security survey

Food security surveys estimate several key proxies of food security, so it is difficult to determine one prevalence a priority to be used in the calculation of the sample size. It is best, therefore, to use 50 percent in the sample size calculations, which gives you the most conservative (largest) sample sizes. On the other hand, a larger margin of error is usually acceptable with food security indicators than with nutrition indicators. For example, a sample size of 200 households with a design effect of 2 gives a 95 percent confidence interval of between 40.5 and 59.5 percent. This accuracy is sufficient for most food security analysis.[48]

Reconciling food security survey and nutrition survey requirements

When reconciling the required sample sizes for nutrition surveys versus food security indicators, the first question to ask is the purpose of the nutrition data. As presented in section 6.2.2, the primary goal of collecting nutrition information in a CFSVA is to analyse the link between food security and nutrition. This can be accomplished even with the smaller sample sizes. However, if the secondary goal of collecting nutrition data is also included in the scope of the survey – to provide accurate and precise prevalences of key nutrition indicators – then the nutrition sample size has to be adequate for that goal.

47. Identified as "common mistake #6" by CDC and WFP, 2005, *A Manual: Measuring and Interpreting Malnutrition and Mortality.*
48. We could also state with 75 percent confidence that the interval is between 44.3 and 55.7 percent. Another argument in WFP's analysis is that we do not need these "scientific" levels of confidence to inform programmes.

In CFSVA practice
If the CFSVA aims only to study the link between food security and nutrition with the nutritional information gathered (using under-5 anthropometry as an example), then the smaller sample size as recommended for food security surveys is acceptable: All households are eligible to be included in the survey. All children under 5 from these households are weighed and measured. This is a representative sample resulting in unbiased, but less precise, population estimates of the anthropometric indicators that, when reported, must indicate the correct confidence intervals.

If the CFSVA needs to make accurate and precise estimates of malnutrition prevalence, larger sample sizes will be needed. Three options are generally considered:

1. Sampling proceeds as described above, but nutrition indicators are reported at a more aggregate level. For example, if a sample of 200 households per "province" is taken, nutrition indicators are reported for a higher aggregation level, a "region," which joins three or four provinces.

2. Sampling proceeds as described above, and in each cluster, additional households are selected, and only anthropometry (and possibly one or two other key indicators) are collected from these households, in order to satisfy the nutrition sampling requirements.

3. The sample size for the household survey is increased to take into consideration the nutritional indicator sample requirements. Then sampling proceeds as above.

4.1.3.3 Rules of thumb: A typical CFSVA sample procedure

Typical sampling methodology for CFSVAs.
Although there are many valid possibilities for taking representative samples in a CFSVA, most countries find similar demands for their sampling. Here is an example of a sampling methodology that follows this "typical" situation.

In Haiti, the WFP country office wanted estimates of food insecurity and other key indicators for the rural population at the department level. (Departments are the main administrative boundaries.) There are 10 departments in Haiti. Additionally, FEWS NET and the Government, both key partners, wanted estimates for the FEWS livelihood zones. There are eight livelihood zones in Haiti, although "urban" was excluded, and "sea salt" included as part of "Plains" under "Monoculture zone" due to its very small size and population. This reduced the zones to six strata. Livelihood zones and departments do not share common boundaries.

The following map shows the departments and livelihood zones in Haiti.

Figure 4.3: Livelihood zones in Haiti

To determine a sample size for any geographic stratum where an estimate will be made in the report, CFSVA sampling typically follows some rules of thumb. These include:

- A two-stage cluster sample is drawn. Usually a minimum of 25 clusters (typically villages) per stratum is desired.
- A minimum sample size of 250 households per stratum is desired, 300 is better, if time and logistics allow. A sample of over 300 is rarely necessary (see table here).
- Between 10 and 15 households are sampled in each cluster; in the case of Haiti, 12 HHs per cluster.
- Typically a design effect of 2 is assumed under these conditions.

Following these criteria, the 95 percent confidence intervals for the different cluster sample sizes are the following:

Sample Size (HHs)	Design Effect	Assumed Prevalence (%)	95% Confidence Interval (%)
200	2	50	40.4–59.6
	2	20	13.3–28.9
250	2	50	41.4–58.6
	2	20	13.9–27.9
300	2	50	42.1–57.9
	2	20	14.4–27.1
400	2	50	43.1–56.9
	2	20	15.0–26.1
500	2	50	43.8–56.2
	2	20	15.5–25.4

As seen in the table, only small decreases in confidence intervals occur when increasing sample size from 250 or 300 to 500.

There may be a need to make a statement about a sub-stratum. This is particularly common post-survey, where a particular group (geographic, livelihood, etc.) of households displays unexpected findings that might be worth highlighting in the report. In the Haiti example, the livelihood zone "Plateau agro-pastoral" was thought to have some possible geographic distributions of livelihoods that would better inform future use of the zones. This implies that, in analysis, this zone may be further split into separate strata. For example, if 30 percent of the zone falls into one area during the analysis, a sample of 250 households in the zone will yield information on only 75 households, making statements about that area imprecise. However, if 500 households were sampled in that zone, 150 households would be included in the survey in that area – which would allow for tighter confidence intervals and therefore more confidence around statements made about that area. In the case of Haiti, it was decided to draw a sample that could provide estimates either at the livelihood zone level, or at the department level.

Although estimates in each of the intersections would be interesting and useful, this would result in 35 strata. If the minimum sample of 250 per stratum were used, this would give (35 strata x 250 HH per stratum =) 8750 HH. This sample size would be far beyond the available resources for the survey.

The next step is to gather relative population estimates. The following population information was calculated. (These numbers reflect population, but households could equally be used.) This information shows the intersections of all departments and livelihood zones.

This data could come from a recent census, other surveys, or Landsat information using GIS techniques to estimate the populations by geographic area.

Table 4.2: Intersections of departments and livelihood zones for sampled rural populations

Rural Population in the Sampling Frame	**Livelihood zone**					
Department Name	Dry agro-pastoral (plus sea salt production)	Plains monoculture	Humid mountain agri.	Plateau agro-pastoral	Agri. semi-humid	Dry agri. and fishing
Artibonite	141 591	267 176	285 739	32 452	n/a	177 205
Centre	n/a	n/a	84 904	405 120	n/a	1 954
Grand'Anse	n/a	n/a	304 313	n/a	47 287	185 229
Nord-Est	69 058	23 306	108 253	n/a	n/a	n/a
Nord-Ouest	246 964	n/a	103 817	n/a	37 455	27 104
Nord	n/a	186 352	355 576	41 094	n/a	n/a
Ouest	41 160	380 199	231 616	n/a	13 319	412 895
Sud-Est	634	n/a	134 433	n/a	68 821	225 526
Sud	20 581	61 542	89 989	n/a	112 206	236 781

If Haiti wanted estimates only for the 10 departments, using 250 HHs per stratum, this would give a total sample size of 2,500 HHs. If only the 6 livelihood zones were considered, a sample size of 1,500 HHs would be needed.

An initial total sample (all strata) was set at about 2,500. This would give about 250 HHs per department, and at least that per livelihood zone. A cap of about 3,000 HHs was set for the entire sample due to logistics and financial limitations. (Note that in this example, Grand'Anse and Nippes were combined, as they did not have separate population figures at the time of the sampling.)

In Excel, a dynamic spreadsheet (Table 4.3) was created to explore possible sample sizes that would provide an appropriate sample size at both department and livelihood zone.

- The population sizes (column C) were taken from Table 4.2.
- The percentage of total population in each line was calculated based on the population size of the line (value in column C cell) divided by total rural population (sum of column C).
- A proportionate sample size (column F) was estimated based on the percentage of total population (e.g. cell F1 = cell E1 x sum of column D).
- Column F was copied (values, not formulas) and placed in column D. (These values were later altered). Column G was calculated using the formula presented in section 4.1.2.8 on weighted data analysis.

Table 4.3: Example of CFSVA sampling procedure

	A	B	C	D	E	F	G
	Department	Livelihood zone	Total population size (rural)	Population of each combination as proportion of the population in entire sampling frame (%)	Planned sample of HHs	Sample size of each combination as proportion of entire sample (%)	Standardized weight (how to duplicate each HH to have a proportionate overall sample) G = D/F
1	Artibonite	Dry agro-pastoral	141 591	2.7	60	2.0	1.38
2	Artibonite	Plains monoculture	267 176	5.2	120	4.0	1.3
3	Artibonite	Humid mountain agri.	285 739	5.5	120	4.0	1.39
4	Artibonite	Plateau agro-pastoral	32 452	0.6	24	0.8	0.79
5	Artibonite	Dry agri. and fishing	177 205	3.4	72	2.4	1.44
6	Centre	Humid mountain agri.	86 858	1.7	48	1.6	1.06
7	Centre	Plateau agro-pastoral	405 120	7.8	300	9.9	0.79
8	Grand'Anse/Nippes	Humid mountain agri.	304 313	5.9	180	6.0	0.99
9	Grand'Anse Nippes	Agro-pastoral semi-humid	47 287	0.9	48	1.6	0.58
10	Grand'Anse/Nippes	Dry agri. and fishing	185 229	3.6	132	4.4	0.82
11	Nord-Est	Dry agro-pastoral	69 058	1.3	96	3.2	0.42
12	Nord-Est	Plains monoculture	23 306	0.5	36	1.2	0.38
13	Nord-Est	Humid mountain agri.	108 253	2.1	120	4.0	0.53
14	Nord-Ouest	Dry agro-pastoral	246 964	4.8	156	5.2	0.93

	A	B	C	D	E	F	G
	Department	Livelihood zone	Total population size (rural)	Population of each combination as proportion of the population in entire sampling frame (%)	Planned sample of HHs	Sample size of each combination as proportion of entire sample (%)	Standardized weight (how to duplicate each HH to have a proportionate overall sample) G = D/F
15	Nord-Ouest	Humid mountain agri.	103 817	2.0	60	2.0	1.01
16	Nord-Ouest	Agro-pastoral semi-humid	37 455	0.7	36	1.2	0.61
17	Nord-Ouest	Dry agri. and fishing	27 104	0.5	24	0.8	0.66
18	Nord	Plains monoculture	186 352	3.6	108	3.6	1.01
19	Nord	Humid mountain agri.	355 576	6.9	180	6.0	1.16
20	Nord	Plateau agro-pastoral	41 094	0.8	24	0.8	1.01
21	Ouest	Dry agro-pastoral	41 160	0.8	24	0.8	1.01
22	Ouest	Plains monoculture	380 199	7.4	180	6.0	1.24
23	Ouest	Humid mountain agri.	231 616	4.5	96	3.2	1.41
24	Ouest	Agro-pastoral semi-humid	13 319	0.3	24	0.8	0.33
25	Ouest	Dry agri. and fishing	413 529	8.0	192	6.3	1.26
26	Sud-Est	Humid mountain agri.	134 433	2.6	72	2.4	1.09
27	Sud-Est	Agro-pastoral semi-humid	68 821	1.3	60	2.0	0.67
28	Sud-Est	Dry ag. and fishing	225 526	4.4	132	4.4	1
29	Sud	Dry agro-pastoral	20 581	0.4	12	0.4	1.01
30	Sud	Plains monoculture	61 542	1.2	36	1.2	1.01
31	Sud	Humid mountain agri.	89 989	1.7	36	1.2	1.46
32	Sud	Agro-pastoral semi-humid	112 206	2.2	84	2.8	0.78
33	Sud	Dry agri. and fishing	236 781	4.6	132	4.4	1.05
		TOTAL	**5 161 651**	**100.0**	**3 024**		

In this example, it is assumed that household size does not change across the rural area. Hence for calculating weights, we use the number of people in a certain area, instead of the number of households, as the basis for the calculations.

Next, the totals from column D were calculated for each department and each livelihood zone using formulas to retain the dynamic nature of the spreadsheet.

	Total sample
Artibonite	396
Centre	348
Grand'Anse/Nippes	**360**
Nord-Est	252
Nord-Ouest	276
Nord	312
Ouest	516
Sud-Est	264
Sud	300

	Total sample
Dry agro-pastoral	348
Plains monoculture	480
Humid mountain agri.	912
Plateau agro-pastoral	**348**
Agro-pastoral semi-humid	252
Dry agri. and fishing	684

Finally, the samples from each department-livelihood zone combination were adjusted following the following criteria:

- All strata (either department or livelihood zone) should have a minimum of 250 HH.
- All sample sizes should be a multiple of 12 (as it was planned to have 12 HH per cluster, so the sub-strata sample sizes would then all give a round number of clusters).
- The total sample size should remain at or below 3,000 households.
- The sampling weights should not become too extreme (it was attempted to keep them all between 0.75 and 1.5, where possible).
- All strata combinations should have a non-zero sample size, and preferably above 36 (if possible).
- The Grand'Anse/Nippes departments, as they are combined in drawing the sample, should be oversampled to the extent possible to allow for later disaggregation between them.
- The livelihood zone Plateau Agro-pastoral should be oversampled, as a secondary analysis of distribution of livelihoods within this livelihood zone was desired.

These adjustments to the sample sizes required some degree of "playing" with the numbers to achieve a sample that satisfied the requirements proposed, and to have a sample that appeared reasonable to the partners involved in the survey. Especially since each "combination" at the same time contributes to the total sample size of a livelihood zone and a department.

Rules of thumb for effective "playing":

- If the totals of some departments or livelihood zones are insufficient:
  - add households/clusters to the intersections where both the department and livelihood zone are lacking in numbers;
  - add households/clusters to the cells of strata with insufficient numbers, add them first where the sample weight is highest.

Once the sample sizes were finalized, the villages (clusters) were selected using PPS selection within each of the department/livelihood combinations, following the guidance for cluster selection.

Within each village, households were selected using simple random sampling, as described in the section on selecting households within clusters.

4.1.4 Key references: Sampling

- Magnani, R. 1997. *Sampling Guide*. Food and Nutrition Technical Assistance Project, Academy for Educational Development, Washington, D.C.
- TANGO International. 2002. *Household Livelihood Security Assessments: A Toolkit for Practitioners.* Prepared for CARE International.
- ibid. 2007. *Monitoring and Evaluation Manual. Prepared for ADRA International.*
- WFP. 2004. *Sampling Guidelines for Vulnerability Analysis: Thematic Guidelines*. Rome.

4.2 HOUSEHOLD DATA COLLECTION

4.2.1 Introduction

Developing questionnaires is a central component of the survey design and implementation process. The quality of any analysis depends on asking the right types of questions and getting reliable answers. At the same time, questionnaires are directly linked to data management and storage. The form a questionnaire takes has direct implications on how the resulting data will be organized for use by an analyst.

While VAM staff may be responsible for different types of assessments with varying objectives, food security remains the central theme. In this context, it is important to ensure that appropriate guidance is available for VAM staff on how best to develop the tools that capture the right information needed for food security analysis.

WFP/VAM surveys use both a quantitative and a qualitative approach in obtaining data. Hence the data collection tools necessarily incorporate components of both approaches.

These guidelines focus on the household questionnaire, as it is the most commonly used data collection tool in a CFSVA survey.

4.2.2 Objective of household data collection

The objective of a household survey is to gather quality indicators, in a standardized way, which after analysis will provide the useful statistics required to fulfil the objectives of a CFSVA, EFSA, or FSMS.

A good-quality questionnaire is a necessary, but not sufficient, tool for obtaining primary data reflecting the real situation. Other necessary tools are: enumerator training, an unbiased sample, and the collaboration of respondents.

The starting point for designing a questionnaire is the list of indicators we want to collect through the survey. The CFSVA guidelines give definitions of standardized indicators collected at the household level. The indicators should be, as much as possible, compatible with generally accepted indicators from other surveys (DHS, MICS, etc.) and other organizations.

4.2.3 Conducting household data collection[49]

Preparing the ground
Pre-survey publicity is essential. Enumerators should not show up unannounced to demand information, as that approach is unlikely to be successful. Letters of introduction should be sent to the appropriate officials, community leaders, etc. These should contain an explanation of the purpose of the research, the procedure for selection, the subject matter to be covered, and an assurance of confidentiality and anonymity.

49. This material is partly extracted from Devereux and Hoddinott, 1992, *Fieldwork in developing countries.*

The first step to a successful interview is properly introducing oneself to the respondent and obtaining informed consent.[50] Enumerators must briefly explain to those being interviewed the purpose of their study, who has funded and supported it, how the data will be collected, the expected duration of the interview, and how the results may be used. If a respondent does not understand the purpose of the interview or does not wish to participate, the survey may end up with inaccurate or misleading answers. The respondents should be aware that they will not receive payment or any other form of compensation, but that their participation is voluntary and that they may refuse to participate in the interview or stop it at any time. Consent must be derived from the actual people involved, not just officials or leaders.

Equally important is talking with an appropriate respondent. The respondent should be an adult member of the household – not a guest – and preferably the household head or her/his spouse. If nobody suitable is available, skip this household and move to the next on the list, returning later to interview the household, if possible.[51]

Interacting with respondents

An essential qualification for successful fieldworkers is a demonstrable and genuine interest in other people. The ideal interview is a friendly conversation between enumerator and respondent. The posing of questions and the noting of replies should have the flow and pattern of a dialogue.[52] This is not accomplished if enumerators are impolite or brusque; nor is it possible if poorly trained enumerators fumble their way through the questionnaire.

A common pitfall is the mechanical recitation of questions without thinking about the responses being given; often this leads to extensive work for those cleaning data during analysis. Time must be taken during the interview. Questions must be asked carefully, making sure that respondents have understood them correctly. It may be necessary to repeat questions and probe to be sure the answer recorded is the one intended.

The interview setting

Interviews are usually one-on-one encounters. When the research deals with personally sensitive matters, the presence of outsiders, or even other family members, may inhibit respondents, embarrassing them into evasion or silence. Onlookers may encourage respondents to answer untruthfully. For example, working during food crises is problematic because respondents have an incentive to understate their stocks of grain and their general wealth, with the expectation that food aid will be brought into the community. This tendency will be exacerbated during a public interview, since a respondent who admits to being wealthy may face demands for help from poorer neighbours or relatives. In general, the more sensitive the topic, the stronger the case for conducting the interview in private.

50. Wilson, 1992.
51. WFP. EFSA Handbook 2009.
52. Casley and Lury (1987:111)

It is critical to be mindful of when the interview is conducted, as the process can be an imposition on respondents' time. Day of the week and time of day are important in both rural and urban settings. Attempt to meet on days and at times convenient to respondents. For example, in urban settings, interviews outside of regular work week hours may be necessary. In rural settings, women may be busy on market days and when preparing meals; men may be busy working in the field at a particular time of day. One way of ensuring that interviews are not an imposition is to make appointments to see people. An interview should never last more than 90 minutes and should be held in a place convenient for the respondent. Urban teams and respondents should be protected from violence and crime by interviewing only at "safe times."

4.2.4 Modules of questions

The questions used to construct single indicators, or different indicators related to similar topics, are usually organized in modules, which should be ordered logically in a questionnaire. Within a module, questions follow a logical flow, and should not be redundant.

Types of modules in a food security household questionnaire

Based on an extensive review of household questionnaires, commonly used modules have been divided into four broad categories:

1. ***Core modules with standard, non-changeable, questions.*** These modules have been tested and used in several CFSVAs, EFSAs, and FSMS and must be used in all food security assessment questionnaires. They contain standard questions (formulated in a standardized way but with country-specific adaptations, e.g. food items, expenditure items) that have proven to be useful for data analysis and answering the five VAM questions. The modules were identified as:

   - Food consumption patterns (including sources of foods consumed)
   - Expenditures
   - Household assets
   - Sources of water
   - Access to sanitation

2. ***Core modules with questions that are flexible (i.e. changeable) depending on specific contexts and survey objectives.*** Some standards have been developed for these modules – but there are slight variations from country to country. These modules are central to WFP food security assessments and essential to answering the five VAM questions. Specifically, the modules containing flexible questions are:

   - Household composition/demography and education
   - Housing materials (walls, floors)
   - Access to credit, indebtedness
   - Livelihoods/sources of income
   - Agriculture
   - Livestock
   - External assistance (food and non-food)
   - Shocks, coping, and/or coping strategies index

3. ***Non-core modules with non-changeable questions.*** These modules are sometimes, but not always included in CFSVAs, EFSAs and FSMS. Global standards have been established for these modules, and therefore the questions should not be adapted or changed. The data are very relevant in terms of answering the five VAM questions. Specifically, this type of module includes:
   - Maternal health and nutrition
   - Child health and nutrition

4. ***Non-core modules with changeable/flexible questions.*** This type of module is sometimes, but not always included in WFP food security assessments. These modules are quite flexible in terms of structure and the types of questions posed. The main modules that fall within this category are:
   - Migration/movement, displacement status
   - Remittances

5. For each one of the four categories and modules contained therein, a series of guidance notes have been written that include the following information:
   - Main purpose of the module
   - Current limitations of the module
   - Creation of the module (i.e., how-to)
   - Modifications that can be made to the module
   - Links to other modules

Each category and associated modules are detailed in section 4.2.4.1.

4.2.4.1 Core modules with standard, non-changeable questions

Module Title: Food Consumption Patterns (including sources of food)

Main purpose of the module: This module allows the analyst to calculate the food consumption score (FCS) for each household and to investigate current food consumption patterns.

Limitations of the module
- The module cannot provide the caloric value or nutritional adequacy of the household diet.
- The module cannot measure the quantity of the food consumed.
- The module does not look at the intra-household differences in consumption.

Creation of the Food Consumption Patterns module
Before recording the dietary diversity and frequency of the household diet, ask the following questions as to the number of meals consumed:
i) Yesterday, how many times did adults eat?
ii) Yesterday, how many times did children 6 to 12 years of age eat (should link with demographic section)?

List the food items belonging to food groups typically eaten in a specific context. The list (which is country specific) should contain: (a) staples and food eaten commonly throughout the study area; and (b) preferred items (e.g. maize versus millet, cassava versus Irish potatoes). The list of food items is country specific and should reflect what is typically consumed in the country. However, food items should be listed in such a way that allows their aggregation into the food groups used for computation of the FCS.

- The list should include between 15 and 20 food items. The list is not meant to be comprehensive of all food items found in the country. Instead, it should reflect the basic items found in a general diet (e.g. oil, salt, meat, dairy products). Use of the word "other" can make the list comprehensive enough.

- Corn-soya blend (CSB) should be considered a separate food item.

- If particular condiments are consumed with staples, it is important to identify them as such and not group them with that food item (e.g. fish powder with fresh fish, milk in tea and glass of milk). Training should be given on whether to include condiments in the analysis.[53]

- It is important to ask about combination food items. For example, when a household indicates that maize and sauce were eaten, and the sauce is prepared with oil, vegetables, salt, and chicken, these items should be indicated as consumed, with the amounts of oil, salt, vegetables, and chicken regarded as more than mere condiments.

- If one member of a household consumes food away from the household, the items eaten should not be recorded. If the entire household ate outside of the household, then the items consumed should be recorded.[54]

- For all food items, the recall period is set at the previous seven days. The purpose is to capture the number of days out of seven that a particular food item was consumed.

- Aside from the food items it is important to identify the primary sources from which the food was acquired (this can be either the primary source or the two primary sources). Generally these sources are: own production; hunting, fishing and gathering;[55] exchange; borrowed; purchased; gift; food aid; and credit.

- All items should have a numeric value. There should be no empty cells. If no consumption is reported, then the source and number of days is recorded as zero.

53. This issue needs further discussion and consensus.
54. This is particularly acute in urban and peri-urban areas.
55. If any of these food sources is a specific and important activity, then hunting, fishing, and gathering can be split up.

Table 4.4: Example of a Food Consumption Patterns module
Yesterday, how many meals did the _____ in this house eat?

Adults Children (< 6 yrs)

Food Item	No. of days eaten over last 7 days	Food source (main and secondary)	Food Source Codes
Maize	|__|	|__|,|__|	1 = Own production (crops, animals)
Rice/paddy	|__|	|__|,|__|	2 = Hunting, fishing
Millets	|__|	|__|,|__|	3 = Gathering
Wheat/Barley and other cereal products	|__|	|__|,|__|	4 = Borrowed
Roots and tubers (potatoes, yam)	|__|	|__|,|__|	5 = Purchased with wages
Pulses/lentils	|__|	|__|,|__|	6 = Exchanged labour for food
Fish	|__|	|__|,|__|	7 = Exchanged items for food
White meat (poultry)	|__|	|__|,|__|	8 = Gift (food) from relatives
Pork	|__|	|__|,|__|	9 = Food aid (NGOs, etc.)
Red meat (goat, sheep)	|__|	|__|,|__|	10 = Other (Specify: _____)
Red meat (buffalo)	|__|	|__|,|__|	
Eggs	|__|	|__|,|__|	
Milk/curd/other dairy products	|__|	|__|,|__|	
Vegetables	|__|	|__|,|__|	
Fresh fruits	|__|	|__|,|__|	
Oil/fats/ghee/butter	|__|	|__|,|__|	
Sugar/sweets	|__|	|__|,|__|	
Salt/spices/condiments	|__|	|__|,|__|	

- The main indicators emanating from the analysis of these data are: (a) number of days out of seven that items and food groups are consumed; (b) household FCS; and (c) percentage contribution of the sources to the household food basket over the previous seven days.
- This module has to be adapted to the context. The food items can be changed, but the exact same eight food groups (staple food, pulses, vegetables, fruits, meat and fish, milk, sugar, oil, condiments) should always be used.[56]

Modifications to the module

- A separate column can be added to the table if an item was consumed in the previous 24 hours. This would be a way to incorporate the Household Dietary Diversity Score (HDDS) indicator in the module. However, the HDDS uses a different data collection methodology and a different questionnaire.[57]
- The number of meals people consumed in the previous 24 hours by age cohort groups in the household (e.g. children under 6, children 6 to 12, and children 13 to 18) can be modified when specific information on child food consumption (or other age cohorts) is needed (e.g. in order to programme a school-age child or MCH food aid intervention).
- Contribution of sources in the previous year and quarters, which can provide information about seasonality for food security analysis, can be modified. Information for a few key staple food groups or for general food consumption may suffice. The source categories should be the same as those used for the food sources. See the following table for an example:

56. See WFP guidelines: *Food Consumption Analysis*, at http://vam.wfp.org/MATERIAL/FCS_Guidance.
57. See FAO, 2007, *Guidelines for measuring household and individual dietary diversity*, June.

In the last calendar year (2006) what was the contribution of (source) to your annual food consumption? How does this differ throughout the year?
(Use proportional piling, or divide the pie method, to estimate the relative contribution of each source to total food consumption)

Food source	Annual (%)	Jan.–March (%)	April–June (%)	July–Sept. (%)	Oct.–Dec. (%)
Own production					
Hunting, fishing, gathering					
Purchases					
Gifts/borrowing					
Food aid					
Total contribution	**100**	**100**	**100**	**100**	**100**

Food Consumption Score (FCS) and Household Dietary Diversity Score (HDDS)

WFP and FAO both use measurements of dietary diversity in their assessments and monitoring systems. WFP has adopted a methodology and tailored it to its own information needs in terms of data collection and analysis of food consumption. FAO uses a methodology based on the Demographic and Health Survey (DHS) procedures developed by FANTA. For both approaches, standard methodologies have been developed to calculate indicators of dietary diversity and consumption frequency.

The Food Consumption Score (FCS). Information is collected from a country-specific list of food items and food groups. The household is asked about the number of times (in days) a given food item was consumed over a recall period of the past seven days. Items are grouped into eight standard food groups (each group has a maximum value of seven days/week). The consumption frequency of each food group is multiplied by an assigned weight based on the nutrient content of a portion. Those values are then summed to obtain the FCS. The FCS has a theoretical range from 0 to 112; WFP has defined thresholds (WFP 2007) to convert the continuous FCS into categories creating three food consumption groups (FCGs): poor, borderline, and acceptable.

The Household Dietary Diversity Score (HDDS). A standard list of 16 food groups, the same for any country/context, is used to gather information on food consumed in the past 24 hours. Information for each group is of a bivariate type (yes/no). To calculate the HDDS, the 16 food groups are aggregated into 12 main groups. All food groups have the same importance (relative weights equal to 1), with each group consumed providing 1 point. The HDDS is the simple sum of the number of consumed food groups (it goes theoretically from 0 to 12). For analytical purposes, the HDDS is often ranked into thirds or quartiles.

Both the FCS and HDDS are used as proxy indicators of household access to food. Data collected for both indicators can also be used to consider dietary patterns and the consumption of specific foods. The FCS and HDDS are used for monitoring economic access to food and surveillance at decentralized levels; moreover, the FCS is used for classifying households who are food insecure, while the HDDS is used for monitoring dietary quality.

Link with other modules

- For internal consistency, it is important that the food items listed in the consumption module be reflected in the expenditures module.

Sources of inspiration

- The standard food consumption module currently adopted for the CFSVAs is unique in its format and methodology. FAO's food consumption module is based on a 24-hour recall. The list of food items in the FAO module is different from WFP's in that it focuses more on diversity and specific food groups (e.g., vitamin A–rich foods).

Module title: Expenditures

Main purpose of the module: This module allows the calculation of household expenditure (in cash)–related indicators. Expenditures are useful as a proxy for wider purchasing power, which is an important component of food access. Moreover, understanding expenditures on specific items allows the analyst to determine how households allocate scarce resources and give priority to competing needs.

Limitations of the module

- Cannot estimate the value of own production (section revolves around only cash expenditures). Collecting consumption and expenditure data is tricky because of the varying extent to which households consume out of their own production, which is not collected in this questionnaire and hence reduces its usefulness.
- Cannot estimate the quantity of food items purchased.
- Food expenditure is linked to the season (e.g. is lower after harvest).
- Non-food expenditures (especially education) are also seasonal.

Creation of the Expenditure module

- List of food and non-food items that are mutually exclusive and yet will encompass all essential expenditures (e.g. education and clothing – do not include school uniforms in both categories).
- List of food items should match those found in the Food Consumption Patterns module, with a few possible exceptions (e.g. collapse meat into one category; include expenditures on condiments).
- In addition to main food items, include "condiments" (e.g. salt, spices, beef-tea cubes, fish powder).
- For all food items, the recall period is set at one month.
- Some non-food items also have a one-month recall period. These are: soap, transport, firewood/charcoal, rent, paraffin, and alcohol/tobacco. (Tip: Collect information on likely daily expenditure for households, not infrequent bulk expenditures.)
- For the remaining non-food items, the recall period is set at six months (prior to the day of survey).
- If other CFSVA and/or EFSAs have been conducted, use the same lists.
- All items should have a numeric value. There should be no empty cells, as that would mean "missing data." If there are no expenditures, use a zero.
- Include debt expenditures (i.e. repayment of loans).

Table 4.5: Example of Expenditure module

	In the past **MONTH**, how much money have you spent on each of the following items or services? (Write 0 if no expenditure.)	**a.** Estimated expenditure in cash	**b.** Estimated expenditure in credit (if applicable)
1	Maize		
2	Wheat/barley		
3	Millet		
4	Rice/paddy		
5	Roots and tubers (potatoes, yams)		
6	Pulses/lentils		
7	Vegetables		
8	Milk/yogurt/milk products		
9	Fresh fruits/nuts		
10	Fish		
11	White meat (poultry)		
12	Pork		
13	Red meat (goat, sheep)		
14	Red meat (buffalo)		
15	Eggs		
16	Oil/butter/ghee (fats)		
17	Sugar/salt		
18	Condiments		
19	Alcohol and tobacco		
20	Soap		
21	Transport		
22	Firewood/charcoal		
23	Kerosene		

In the past **6 MONTHS** (semester), how much money (in cash) have you spent on each of the following items or services?
Write 0 if no expenditure.

24	Equipment, tools, seeds	30	Celebrations, social events, funerals, weddings
25	Hiring labour	31	Fines/taxes
26	Medical expenses, health care	32	Debts
27	Education, school fees	33	Construction, house repair
28	Clothing, shoes	34	Other long-term expenditure (Specify: _____)
29	Veterinary expenses		

The typical indicators emanating from this module include:

(a) total household expenditures (food and non-food);
(b) total per capita expenditures;
(c) percentage food expenditures;
(d) percentage non-food expenditures; and
(e) percentage individual food and non-food items.

Modifications to the module

- Recall period for non-food expenditures can be modified to one year.
- Recall period for education can be realigned with the calendar of payments; however, a conversion of the expenditures to "yearly expenditures" must be allowed.
- Recall period for food expenditures can be reduced (1 week) during an EFSA.
- FSMS may modify the non-food recall period to correspond to the last data collection round, as long as a pro-rated amount can be calculated.
- A credit or exchange/barter column could be added when appropriate, but this will require additional time for survey administering. Based on the experience with previous CFSVAs, such data add little value for food security analysis.
- Pastoralists: include information on veterinary fees, water costs for animals, livestock purchases (this could be placed in the Water/sanitation module, as long as both modules have the same recall period).
- Seasonality of expenditures — information on celebrations, school fees, agricultural inputs, seasonal disease outbreaks (e.g., malaria) and health —could be collected in a community/key informant questionnaire.
- Rent expenditures should be included in the monthly expenditures, unless specific questions on house ownership and rental are posed under the "Housing" section.

Links with other modules

- Questions on expenditures can be asked in other modules, as long as they are not duplicated.
- Link to income sources, food consumption modules (triangulation at field level as well as analysis stages).
- When deciding on a recall period, make sure that all relevant secondary data are reviewed to ensure that the proper period is chosen. Harmonize all recall periods in other parts of the questionnaire, otherwise enumerators will likely make mistakes with the period used.

Module Title: Household Assets

Main purpose of the module: This module allows the analyst to calculate proxy indicators of wealth and qualify the type of assets the household possesses.

Limitations of the module

- It does not allow for an exact measure of wealth.
- The list is a finite number of wealth and production assets. The selected assets must be typical for the context but allow for inter-household differences to be captured.
- It is advisable that the list contain between 10 and 15 non-perishable assets.

Creation of the Assets module

- Prepare a list of productive and non-productive assets. Guidance on the applicability for the specific country or region can be sought from previous WFP surveys or HEA, MICS, and DHS studies. In the development of this module, questionnaire designers should look at other household surveys done in the same country, especially large-scale government surveys or DHS, and make the asset module of the CFSVA survey compatible with these. A big advantage to this is the ability to compare levels and distributions of assets to determine, for example, if households in your survey are richer or poorer than households found in other data sets.

- Based on the context, assets should include agricultural and journeyman's tools (e.g., pesticide sprayer, plough, mason's tools, carpenter's tools), as these are examples of productive assets.
- "Luxury" assets are a reflection of the wealth and/or social status of a household. They can be used to generate income, though this is not their primary use. Examples are context specific, but can include a radio, television, satellite receiver, mobile phone, car, and microwave.
- Typical household assets should also be included and, though context-specific, can include: a mattress, a lantern, a mosquito net, or a manufactured cooking stove.
- Once the list has been created and verified by local experts as being applicable to the context, list the items by category (basic, productive, and wealth).
- To the right of the "Assets" column, place either 1 (yes) or 0 (no); the asset the household owns can also be circled.

Table 4.6: Example of Household Assets module

Indicator	Productive/transport assets	
Does your household own any of the following assets? ***Circle all that apply***	1 Shovel/spade 3 Sickle 5 Fish net 7 Rice mill (fuel) 9 Motorcycle 11 Bicycle	2 Plough 4 Weaving tool 6 Pounding mill (wood), foot or hand 8 Rice mill (electricity) 10 Hand tractor 12 Boat/canoe
	Household assets	
	13 Sleeping mats 15 Table 17 Stove (gas/fuel) 19 Generator (run by fuel)	14 Bed 16 Radio 18 Generator (run by water) 20 Mosquito net

- The main indicators emanating from this module are: (a) percentage of households owning an asset (e.g. radio), and (b) wealth index.
- Enumerators can be given guidance on excluding assets beyond repair.

Modification to the module

- The number of each asset can be recorded (e.g. 4 chairs, 2 beds, 6 radios), but this adds little value for food security analysis.

Link with other modules

- There should be an internal consistency between productive assets and household economic activities.

Source of inspiration

- Household questionnaires from the DHS always include comparable questions regarding the ownership of productive and non-productive assets. Similar to DHS surveys, CFSVAs look at assets ownership and adopt the same methodology to compute the wealth index (i.e. Principal Component Analysis).

Module Title: Sources of Water and Sanitation

Main purpose of the module: This module allows the analyst to estimate the percentage of the population using improved drinking water sources and improved sanitation, which is a commonly used indicator to assess hygiene at the household level.

Limitations of the module

- The module cannot tell us about the quality of the drinking water for each type of source, the quantity of water the household is drinking, or the household water storage practices.
- The module alone cannot tell us the impact of poor hygiene on food security and nutrition. The link between poor water access and hygienic conditions, malnutrition and food security has to be statistically explored and proven. Appropriate expertise is required for this.

Creation of the Water and Sanitation module

- Based on the accepted United Nations guidance, the primary household water sources must be grouped accordingly: water tap in household; water tap in compound; public stem pipe; borehole; protected dug well; protected spring; rainwater collection; unprotected well; unprotected spring, river, or pond; vendor provided; tanker truck; bottled water. The list is context-specific. An example of the question follows:

What is the main source of drinking water for your household? *Circle one*	1	Piped water in-/outside	5	Mountain source (incl. gravity water feeder system)
	2	Well/borehole protected	6	Rainwater from tank
	3	Well/borehole unprotected		
	4	River, stream, or dam	7	Other

- Based on the accepted United Nations guidance, household sanitation sources must be grouped as follows: flush toilet, pail flush, simple pit latrine, ventilated improved pit latrine, public/shared latrine, open pit latrine, bucket latrine, bush. An example of the question follows here:

Where do members of your household normally go to the toilet? *(Do not read answers aloud.)* ***Circle one***	1	Flush latrine/toilet with water	4	Communal latrine
	2	Traditional pit latrine (no water)	5	None/bush (go into forest)
	3	(Partly) open pit (no roof or no wall)		

- The main indicators from these modules are: (a) percentage of households using improved drinking water sources; and (b) percentage using improved sanitation.

Modifications to the module

The following questions can be added to further the understanding of household access to water:

- How much time is required to collect water (round trip)? This can either be a specific time (e.g. 35 minutes) or a categorical response (e.g. 1 to 3 hours).
- What is the distance from the water source (one way; usually collected as continuous variable in kilometres)? This information can be used for spatial analysis of accessibility of resources. Appropriate skills and other relevant data must be available for this type of analysis.
- Who collects the water (e.g. girls, boys, women, men)? This information can be collected in the Key Informant or Community questionnaire.
- Do you pay for water? How much do you pay for the water per day/week/month (depending on the context)?
- What is the seasonal variation of your water source? Is there is a second source?

Links with other modules

- If questions are asked regarding expenditure on water, they should not be repeated in the Expenditure module, and the reference period should be the same (e.g. 1 month).
- The link between water source and sanitation and the nutritional status of the households can be explored.
- Components of a wealth index.

Sources of inspiration

- Questions on water sources are always included also in the MICS and DHS questionnaires. The main differences between CFSVA and DHS/MICS modules are:
  i) CFSVAs focus on the source of drinking water, whereas DHS and MICS consider sources of water used for cooking and hand-washing.
  ii) A more detailed list of options is typically included in the DHS/MICS questionnaires than in the CFSVA questionnaires.
- Questions on sanitation are always included in the MICS and DHS questionnaires. In DHS and MICS, the options for type of toilet are usually more detailed. CFSVA HH questionnaires do not include questions on the number of people sharing toilet facilities.

4.2.4.2 Core modules with flexible questions

Module Title: Demography and Education/Household Composition Roster

Main purpose of the module: This module records demographic characteristics of households.

Limitations of the module

- A household is usually defined as people living in the same compound and eating "from the same pot," forming a clear socio-economic entity. The head of the household makes the major decisions. However, these definitions have to be adjusted to be in line with country-specific definitions. The definition of a household can bias results (e.g. polygamy can be incompatible with the standard definition of a household). Often the definition of the "head of household" is culturally defined and may not reflect "who is making the key decisions" or who is the bread winner.

- With a comprehensive roster (at the individual level), the module can become quite weighty and time-consuming.
- It will not provide rates of fertility or fecundity.
- Mortality and morbidity rates can be calculated provided that the module is organized in a suitable way (using roster type) and that appropriate questions are asked. However, the usual sampling design, while suitable for food security indicators, can give rather imprecise estimates of mortality and morbidity.
- This module is not meant to capture members of the household who have migrated.

Creation of the Demographics module

- There are two standardized methods for collecting household information. However, depending on the information desired, one approach will be recommended (e.g. mortality; non-age-bound population estimates will require a more detailed enumeration of the household, and a roster approach is the recommended tool). The type of tool will affect the amount of time required to administer and the level of skills required by the enumerators.

As with other modules, the simplest approach is presented as standard base for CFSVA surveys. It can provide a limited but sufficient amount of information. As more information is required, more questions can be added, but the module's structure then becomes more complex, as with the individual roster.

The basic standard: Household Summary module

The following questions are required to capture the household composition and child enrolment:

1. Sex of the head of the household (Male = 1 Female = 2)
2. Age of the head of the household (|__|__| years)
3. Marital status of the head of the household (1 = Married, 2 = Cohabitating, 3 = Divorced/Separated, 4 = Widow/Widower, 5 = Never married)[58]
4. Can the head of the household and the spouse of the head of the household read/write a simple message? (yes/no)
5. Age pyramid: Create a three-column table as here:

Please complete this household's demographics table on the right. This is to record the number of individuals in each age category. Make sure to differentiate between males and females.

Age	Male	Female
a. 0–5 years	|__|__|	|__|__|
b. 6–14 years	|__|__|	|__|__|
c. 15–59 years	|__|__|	|__|__|
d. 60 years or older	|__|__|	|__|__|

These are the basic categories used to calculate the dependency ratio (see "Modifications to the module" for further discussion on adaptation to the local context). There should be no blank cells, if there are no individuals in the age cohort and sex, then a zero must be used.

6. Attendance rate and absenteeism from school: the basic information is collected for primary school children only.

58. Issues of polygamy need to be considered.

Table 4.7: Example of household questionnaire for primary school attendance

		Male	Female
1.10	Number of children attending primary school (6–11 years)?	___	___
1.11	Did anyone miss school for at least 1 month in the last year?	1 Yes	2 No → go to 1.13

		Male Children	Female Children
1.12	If yes, why? **CIRCLE THE MOST IMPORTANT REASON**	1 Sickness 2 Work for money or food 3 Domestic work (gardening, fetching water) 4 Taking care of siblings 5 School is far away/located in insecure area 6 No money for school fees/ school costs 7 Refused to go 8 Other (Specify:_____)	1 Sickness 2 Work for money or food 3 Domestic work (gardening, fetching water) 4 Taking care of siblings 5 School is far away/located in insecure area 6 No money for school fees/ school costs 7 Refused to go 8 Other (Specify:_____)
1.13	If there are boys, girls, or both who do not attend school, what is the main reason? **CIRCLE THE MOST IMPORTANT REASON**	1 Sickness 2 Work for money or food 3 Domestic work (gardening, fetching water) 4 Taking care of siblings 5 School is far away/located in insecure area 6 No money for school fees/ school costs 7 Refused to go 8 Other (Specify:_____)	1 Sickness 2 Work for money or food 3 Domestic work (gardening, fetching water) 4 Taking care of siblings 5 School is far away/located in insecure area 6 No money for school fees/ school costs 7 Refused to go 8 Other (Specify:_____)

The main indicators created through the household roster/summary are:

- Average household size;
- Percentage of male- and female-headed households;
- Average age of head of household (aggregated by sex);
- Marital status;
- Age pyramid;
- Literacy rates of household heads and spouses;
- Attendance rates (check definition) and causes for not attending school;
- Absenteeism and causes; and
- Percentage of dependents/dependency ratio.

Modifications to the module

Household Summary

- Total years or completed level of education of household head and spouse.
- The generic child age cohort can vary depending on the primary school age category in the country. For example, in the 2006 Lao PDR CFSVA, the generic school age cohort was split in two (6 to 11 and 12 to 14 years) in order to reflect the school system of the country and the types of information WFP needed for its school feeding programmes.

- Number of days of absenteeism from school in a certain reference time (the two parameters can vary according to local agreement).
- Child labour and adult labour to household economic unit.
- Chronic illness.
- Disabilities.

The decision to move from household summary questions on demography and education to a household roster type of table, where each question is asked for each individual in the household, is usually driven by the need to obtain more articulated information about education, labour, chronic illness, and disabilities for both adults and children. The household roster provides information and statistics at the individual level. This adds valuable information but also difficulties: the administration of the questionnaire is time-consuming, and this method adds data management and analytical difficulties. Again, the choice of having household- or individual-level indicators must be driven by programme need, analytical capacity, and the intended use of this more detailed information.

Household Composition Roster (example in Table 4.8)
Create a table with the following columns, from left to right:

1. Household member code: unique number that must be consistently employed when the table spreads over several pages (e.g. the same individual in the households has the same household member code).
2. Name of the individual (this is usually not entered in the database but is used during the interview).
3. Gender of the individual.
4. Relationship of the individual to the head of the household (head of the household, spouse, child, orphan, uncle, other).
5. Age of the individual in years (never record the months of a child under the age of 1; use a zero). Where age is not known, an event calendar or other estimation tool should be provided.

Adult Cohort

6. Can the household member read/write a simple message (use DHS/MICS definition)?
7. What level of education does the individual have (this should follow the formal schooling system of the country)?

Child Cohort (age groups will depend on the country)

8. Does the child go to school? (0 = Does not go the school, 1 = primary, 2 = secondary, 3 = university; – this is only for school-age children).
9. If not, why?
10. Did the child miss school for at least five days or more in the past month?
11. If yes, why?

Table 4.8: Example of household composition roster

	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.10
Household Member code	**First Name**	**Gender**	**Relationship to head**	**Age in years**	**Marital Status**	**Current level of education**	**Schooling status of children 6-14**	*If attending school*	
	Do not record full name, only a first name to refer to the household member	1 = Male 2 = Female	1 = Head 2 = Spouse 3 = Child 4 = Parent 5 = Sibling 6 = Grandchild 7 = Grandparent 8 = Orphan being taken care of 9 = Other relative 10 = No relation	For children < 12 months, write 0	1 = Married 2 = Cohabitating (not married) 3 = Divorced 4 = Living apart, not divorced 5 = Widow or widower 6 = Not married	1 = No schooling 2 = Some primary 3 = Completed primary 4 = Some secondary 5 = Completed secondary 6 = Vocational 7 = Some university 8 = Completed university 9 = N/A	1 = Attends primary 2 = Attends secondary 3 = Not attending school	Did [name] miss school for at least 1 month in the last year? 1 = Yes 2 = No	What was the reason for missing? 1 = Sickness 2 = Work 3 = Household Work 4 = Taking care of siblings 5 = Long distance to school 6 = School fees not paid 7 = Insecurity 8 = Refused to go
01									
02									
03									
...									
...									
14									
15									

Additional modifications to the Household Composition Roster

- For mortality/morbidity refer to WFP nutrition guidelines. This addition should be made only if required and strongly supported and properly undertaken.
- Child labour and adult labour to household economic unit.
- Chronic illness.
- OVC issues (see HIV/AIDS guidelines).
- Disabilities.

Links with other modules

- Links with the maternal and child modules (same number of under-5s and women listed as in the modules; carry over the codes.
- The age cohorts need to be identical to those used for schooling of children.

Sources of inspiration

- Information on household demographics is typically collected by DHS, MICS, and LSMS questionnaires through a household roster (individual-level information). CFSVAs use either individual level rosters or household-level questions.
- When a roster is included in the CFSVAs, it usually covers the same areas addressed by MICS, DHS and LSMS, including sex, age, position of the HH members, health conditions, educational level, school enrolment/attendance (for children), questions related to OVC. In general, questions on health status and schooling are more detailed in MICS, DHS, and LSMS; in particular, MICS (used by UNICEF to assist countries in filling data gaps for monitoring the situation of children) systematically collects information on OVC and child labour.

Module Title: Housing Materials

Main purpose of the module: The construction materials used in a household are very basic indicators of living standards. They provide analysts with information on a household's standard of living that goes beyond consumption expenditures. Usually materials for floor, roof and walls are recorded, as per example in Table 4.9.

Limitations of the module

- Housing materials provide only an indirect measurement of wealth.
- Certain "luxury" materials might not have been available in the local context, thus preventing the use of these indicators to identify wealthy households.

Table 4.9: Creation of Housing Materials module

3.5	*What is the major construction material of the exterior walls?* **IF POSSIBLE, DO NOT ASK; ANSWER BASED ON YOUR OBSERVATIONS**	1 Concrete/burned bricks 2 Mud blocks 3 Mud and straw	4 Wood 5 Plastic shelter 6 Other (Specify: _____)
3.6	*What is the major material of the roof?* **IF POSSIBLE, DO NOT ASK; ANSWER BASED ON YOUR OBSERVATIONS**	1 Concrete 2 Tiles 3 Straw (grass, papyrus, banana fibres)	4 Wood 5 Plastic shelter 6 Galvanized iron 7 Other (Specify: _____)
3.7	*What is the major material of the floor?* **IF POSSIBLE, DO NOT ASK; ANSWER BASED ON YOUR OBSERVATIONS**	1 Concrete 2 Mud 3 Straw 4 Wood	5 Plastic sheeting 6 Tiles 7 Other (Specify: _____)

Modifications to the module

Additional elements can be included in the Housing Materials module other than construction materials:

- The type of dwelling (single family house, separate apartment, mud house, shelter, other).
- The number of rooms.
- Availability of electricity.

Links with other modules

- Water and sanitation facilities and assets to construct the wealth index.
- Expenditure to cross-check wealth status.

Sources of inspiration

- Information about the main material of the dwelling, floor, roof, and walls is collected in the MICS, DHS, and LSMS questionnaires. As with the DHS, the CFSVA takes into consideration this information while selecting the variables for the wealth index.

Module Title: Access to Credit

Type of questionnaire: The following presents the Access to Credit module as collected in a quantitative/household survey. However, access to credit can be deeply explored in focus group discussions or in community interviews. This would save time and would make the household-level module optional. More information on qualitative tools can be found in Section 5, "Qualitative and Community-level Data in CFSVA."

Main purpose of the module: The module provides for an estimation of household's access to sources of credit and their actual use of credit.

Limitations of the module

- The module does not aim to estimate the amount actually borrowed.
- There is no information collected with regard to the "interest rates" charged to borrowers or other credit conditions.

Table 4.10: Creation of Access to Credit module

3.5	Do you have **access** to a place to borrow money? **Circle all that apply**	1 Yes – relatives/friends 2 Yes – charities/NGOs 3 Yes – local lender 4 Yes – bank	5 Yes – cooperatives 6 Yes – village head 7 Yes – company/middle men 8 No access to credit
3.6	In the last 3 months, how **often** did you use credit or borrow money to purchase food? **Circle one**	1 = Never 2 = On one occasion 3 = On two occasions	4 = On three occasions 5 = On more than three occasions

Modifications to the module

- The first question can be broken into two: Do you have access to credit? If yes, where?
- Additional questions can be added to explore the average amount of debt in addition to the issue of access to credit sources. Experience from past surveys indicates that piece of information might not be fully reliable, perhaps because of people's reluctance to declare their financial status.

Links with other modules

- Expenditure module: if the household manages to pay back their debt.
- Information about access to credit facilities can be gathered in focus groups or community interviews.

Source of inspiration

- Questions on credit usually are not included in DHS and MICS questionnaires, but they are frequently inserted in the LSMS questionnaires, which collect this information at the individual level. Source, frequency, and time needed for reimbursement are also addressed by the LSMS module.

Module Title: Livelihoods and Economic Activities

Main purpose of the module: This module attempts to identify the activities and combinations of activities that sustain households, and their relative importance to a household's income strategy.

Limitations of the module

- This is not a comprehensive livelihood analysis, which includes but is not limited to economic activities. Its main goal is to identify and group households based on a common set of economic activities and their relative importance for risk analysis.
- If absolute values are collected from the economic activities, the sum of those values should not be considered as an income level for the household. This derived income is not intended for poverty analysis.

Creation of the Livelihoods and Economic Activities module

- Prepare a list of economic activities that households would undertake or the main income sources a household would exploit to earn cash or acquire food or services. The list of activities should be based on secondary data and local expert knowledge. It is important to include atypical sources that vulnerable households would exploit to sustain themselves.
- If another CFSVA and/or other EFSAs have been conducted previously, review the activities listed and include (1) those reported in the previous study, and (2) through review, those that might have been excluded. The aim is to differentiate households and minimize the reporting of undefined "other" activities, which are difficult to interpret and could confound results.
- Include a column where households are asked, using proportional piling, to estimate the relative importance of the activities to contributing to the household's income, food, and access to services.
- The module is not meant to be exhaustive in identifying all the activities undertaken by each household. Instead, it is critical to identify the three or four essential activities.
- It is likely that the three or four activities cover almost all income sources of the household. The sum of the three to four contributions should equal 100 percent.[59]

59. When there are more than three to four activities, it must be made clear to the enumerator that the proportions reported are valid only for the identified activities.

- The categories should not be duplicated. For example, if men undertake one type of agricultural activity and women undertake another type, the two activities should be grouped, as the level of analysis is the household.
- The main indicators emanating from this module are: (a) main economic livelihood activities; and (b) percentage contribution of main economic/livelihood activities to household income.

The *minimum* information required can be obtained through one of the following tables:

Table 4.11a: Recommended layout of economic livelihoods table

Activity(ies) undertaken to earn cash or acquire food or services	(√)	Using proportional piling or "divide the pie" methods, estimate the relative contribution to total income of each activity (%)
1 = Agriculture and sales of crops	I__I	I__II__II__I
2 = Livestock and sales of animals	I__I	I__II__II__I
3 = Brewing	I__I	I__II__II__I
4 = Fishing	I__I	I__II__II__I
5 = Unskilled wage labour	I__I	I__II__II__I
6 = Skilled labour	I__I	I__II__II__I
7 = Handicrafts/artisanal work	I__I	I__II__II__I
8 = Use of natural resources (firewood, charcoal, bricks, grass, wild foods, honey, etc.)	I__I	I__II__II__I
9 = Petty trading	I__I	I__II__II__I
10 = Seller, commercial activity	I__I	I__II__II__I
11 = Remittances	I__I	I__II__II__I
12 = Salaries, wages (employees)	I__I	I__II__II__I
13 = Begging, assistance	I__I	I__II__II__I
14 = Government allowance (pension, disability benefit)	I__I	I__II__II__I
15 = Others (Specify:_____)	I__I	I__II__II__I
TOTAL	**1 0 0**	

Table 4.11b: Alternative layout of economic livelihoods table

Activity	**a.** List, in order of importance, household income activities? (Use activity code from a list like the one in Table 4.11a)	**b.** Using proportional piling or "divide the pie" methods, estimate the relative contribution to total income of each activity (%)
Main	I__I	I__II__II__I
Second	I__I	I__II__II__I
Third	I__I	I__II__II__I
Fourth	I__I	I__II__II__I

The advantage of Table 4.11a is that households can list as many activities as they want. Additionally, the output variables obtained from such a table during the data entry process are ready for analysis (see Annex 16). However, some data management has to be done to come up with the main activities at the household level (percentage of household undertaking agriculture, trading, etc.).

The second example presents the question in an easier way, both for the interviewer and interviewee. The interviewer asks what the main (or second, etc.) household's activity is and the interviewee is free to recall without having to reply yes/no to a long list of activities. If data are collected this way, it is easier to calculate percentages of households undertaking determinate activities as their main one. On the other hand, calculating "Contribution from the different livelihood activities" requires more data management skills.

The preference for one module option over the other should depend on the main indicator(s) needed and on the available analytical capacities.

Modifications to the module

- The table can also be expanded to include information on who undertakes or is the key actor in the activity (see above Tables 4.11a and 4.11b).
- The recall period for the combinations of the activities is typically one year. However, depending on the context (e.g., following a rapid EFSA where the period of time it takes household to adapt, and how they adapt, are relevant), the recall period can be reduced.
- The seasonality of activities can be included to identify when key economic activities are undertaken. This can also be done in a community questionnaire, community focus group, or key informant interview.
- Instead of getting relative contributions (percentage), the absolute cash value of the activity can be captured by recording either (1) the estimated value provided by the household; or (2) the provided value within a series of categorical variables. Even though value ranges are commonly more easily collected, categorical variables present more limitations during the analysis phase. On the other hand, the feasibility of collecting truly reliable absolute cash values has to be explored and tested.
- The respondent could also be asked to estimate the percentage of results/goods from each activity that is directly consumed by the household. This question is used to estimate the relative importance of self-production that is directly consumed and is not captured by expenditure indicators. However, this concept has been reported to be difficult to explain both to enumerators and to interviewees; and the analysis is quite complicated and is based on the assumption that a household's total income can be measured through total cash expenditures plus the value of own-produced and consumed goods.

An example of possible modification of Table 4.11b is shown in Table 4.12.

Table 4.12: Modified Livelihoods/Economic Activities module

Using the "Income activity" and "Participant" codes provided below, complete the following table, completely filling in the information for one activity at a time (to earn cash or acquire food or services)

		a. What is your household's [rank] income activity? (Use "income activity" codes)	**a.** What is your household's [rank] income activity? (Use "income activity" codes)	**c.** Using proportional piling or "divide the pie" methods, estimate the relative contribution to total income of each activity (%)
1.	**Main**	|__|__|	|__|	|__||__||__| %
2.	Second	|__|__|	|__|	|__||__||__| %
3.	Third	|__|__|	|__|	|__||__||__| %
4.	Fourth	|__|__|	|__|	|__||__||__| %

Income activity codes

1 = Agriculture and sales of crops
2 = Livestock and sales of animals
3 = Brewing
4 = Fishing
5 = Unskilled wage labour
6 = Skilled labour
7 = Handicrafts/artisanal work
8 = Use of natural resources (firewood, charcoal, bricks, grass, wild foods, honey, etc.)
9 = Petty trading
10 = Seller, commercial activity
11 = Remittances
12 = Salaries, wages (employees)
13 = Porter
14 = Begging, assistance
15 = Government allowance (pension, disability benefit)
16 = Others (Specify:_____)

Participant codes

1 = Head of the household only
2 = Spouse of the head of the household only
3 = Men only
4 = Women only
5 = Adults only
6 = Children only
7 = Women and children
8 = Men and children
9 = Everybody

Links with other modules

- The income activities (e.g. agriculture and livestock) should agree with the households' responses regarding agriculture and pastoral activities.
- Link to expenditure and credit (the total of cash income and credit should correspond with total cash expenditures, which should be verified during data collection and analysis).

Module Title: Agriculture

Main purpose of the module: This module aims to gather information on the practice of agriculture at the household level. In most developing countries, not only is agriculture one of the main income-generating activities of a household, but the majority of the population also practice it. Furthermore, many households, especially in rural areas, produce at least part of the food they consume, through agriculture or home gardening.

The CFSVA should also identify net food sellers, especially in prime agricultural areas, to help programmes aimed at developing local agriculture through local purchases.

Limitations of the module

- Even though food security surveys almost always collect agriculture information, they are different from an agriculture extension survey or a crop and food supply assessment mission (CFSAM).

- In a food security survey, the aim is not to precisely measure the size of cultivated land and yields, but rather to cross-tabulate agriculture-related data with other socio-economic characteristics such as family size, data on income, expenditure, and consumption, for the purpose of presenting a better picture of the livelihood of rural households and to identify possible factors of food insecurity and inform response options later on.

Creation of the Agriculture module

Information obtained from a basic module on agriculture includes:

- Percentage of households having access to land
- Most common types/methods of land access
- Percentage of households cultivating land
- Common crops cultivated
- Prevalence of kitchen gardens

Modifications to the module

- The nature and scope of the study will determine the level of information sought from the module. Hence, the above list is not meant to be comprehensive.
- Additional questions for the key staples: duration of production for own consumption, share of the production sold, share of the production consumed.
- Additional questions for consideration: size of land, major crops grown, source of seeds, use of agrochemicals, source and extent of irrigation and average yields per harvest. Of course, the more information collected the more complex the module becomes for both the enumerator and for the interviewee. Carefully evaluate the needs of the survey in order to avoid overloading the questionnaire with questions that will not be analysed.
- Questions regarding agriculture can be very specific and detailed, or more general covering the agriculture sector as a whole.

Links with other modules

- Livelihood and economic activities
- Productive assets
- Livestock
- Sources of consumed food
- Expenditure

Module Title: Livestock

Main purpose of the module: to gather information on livestock ownership. Livestock can be seen as assets or as main livelihood activities for pastoralists and nomads, but also involves specific vulnerabilities.

Limitation of the module

- Although food security surveys collect livestock ownership information, they are different from livestock surveys.

Creation of the Livestock module

Usually a filter question opens the module. The list of commonly owned animals and wealth status livestock follows.

4.20 -	Does your household own any farm animal(s)?		1	Yes	2	No → next Section
4.21 -	If yes, how many of each of the following animals do you own? (write 00 if none)					
a	Chicken	___	g	Pigs	___	
b	Ducks	___	h	Bulls	___	
c	Other birds	___	k	Cows	___	
d	Rabbits	___	l	Oxen	___	
e	Goats	___	m	Donkey	___	
f	Sheep	___	n	Camels	___	

Modifications to the module

- The filter question can be removed. When the household does not own any livestock, it is important to enter a zero (0).
- The data collection can be simplified by recording single-species ownership only as a categorical bivariate (yes/no).
- However, in appropriate countries, collecting the number of animals allows for computing the synthetic indicator "Livestock Index" through the use of Livestock Tropical Unit values.
- Extra information on amount of fodder needed, average fodder price, and monthly expenditure on fodder can be useful.

Links with other modules

- Livelihood and economic activities
- Agriculture
- Sources of consumed food
- Expenditure

Module Title: External Assistance (food and non-food)

Main purpose of the module: Any external assistance going on in the surveyed area or in the country should be recorded and taken into account when evaluating households' self-reliance and food security.

Limitation of the module

- The module is not designed to record tonnage or quantities of aid delivered/received by each household.

Creation of the External Assistance module

Usually the module is introduced with a filter question. After that, there is the list of food aid programmes, the organizations that provide non-food assistance in the area, and the types of assistance received by the household. The two lists must be customized according to the local context.

Table 4.13: Example of External Assistance module

Has any member of your household received food aid in the last 6 months?	1	Yes	2	No
If yes, please specify the type of programme and the number of beneficiaries in your household? Circle all that apply and specify number of beneficiaries in the last column.	1 2 3 4	School feeding Food for work/food for assets Supplementary feeding Other (Specify: _____)		|__|__| |__|__| |__|__| |__|__|
Has any member of your household received any other type of external assistance besides food aid in the last 6 months?	1	Yes	2	No
If yes, from whom? Circle all that apply	1 2 3 4 5 6 7 8	World Food Programme SAPPROSC/DEPROSC Save the Children UNICEF GT2/SNV/DFID French Cooperation The Government Other (Specify: _____)		
If yes, what type of assistance? Circle all that apply	1 2 3 4 5 6 7	Food products Money allowances/loans For education (fees, books, uniforms) For medical services Construction material, building Agricultural assistance (tools/seeds) Other (Specify:_____)		

Modification to the module

- The list of assistance programmes and organizations has to be context-specific.

Link with other modules

- Livelihood activities – since assistance programmes sometimes focus on certain livelihood activities, there should be a correspondence; for instance, if the household benefits from agricultural assistance, we could expect crop production to be mentioned as a livelihood activity.

Source of inspiration

- Questions on external assistance are usually not included in DHS and MICS questionnaires, but can be found in LSMS questionnaires.

Module Title: Shocks and Coping

Main purpose of the module: This module aims to identify shocks that, in the recent past, have affected the household's ability to acquire food or cash perceived as important by each household, and the types of coping mechanisms used. This will determine which households are prone to be affected by shocks and which have poor coping capacity.

Limitations of the module

- Problem with the definition of shock: A shock is an abnormal event affecting a household's economic status and capacity to feed themselves. Sometimes events are reported as shocks that do not have real consequences on a household's status because they are (or should be considered) normal events (e.g. lack of rain in a desert area). Information on shocks can be more appropriately collected through

secondary data. The problem of shock definition should be carefully considered during questionnaire design and enumerator training.
- Shocks and coping strategies depend on household perception.
- Households may not be able to attribute their coping mechanisms or the consequences of a shock to a particular event.
- Neither intensity of shock nor coping mechanism is measured.
- This is not equivalent to the Coping Strategy Index. If this is a desired output, it should be included as a separate section.
- If an initial filter question is used (e.g. Have you experienced a shock in the last 12 months?) as a skip question, this may lead to unwarranted non-response by households, as some may not understand what is meant by the term shock (i.e. they might not consider specific shocks, such as fire, drought, or war).
- If no initial filter question is used, and each shock is enumerated (i.e. Have you experienced drought? Have you experienced floods?), this may increase the number of responses, as each may be a leading question.

Creation of the Shocks and Coping Strategies module
- The focus of this module is on **shocks that affect the economy of the household or the ability to acquire (produce/purchase) food.**
- The **minimum structure** of this module should be a list of shocks experienced and coping mechanisms used.
- The recommended standard recall period is one year. There are exceptions to this; see "Modifications to the module."
- The module should reflect:
  - Covariate shocks experienced during the period of recall (e.g. one year), including economic and environmental shocks;
  - Idiosyncratic shocks that likely affected the household;
  - Coping mechanisms commonly employed by households within the context, especially those used by known vulnerable households.
    - Coping mechanisms should reflect food and non-food responses (i.e. the list should come from the secondary data review or previous surveys/studies).

Modifications to the module
- Once the shock and coping strategy lists are made, the following options can be added (there is no current agreed-upon standard):
  - A filter question at the beginning of the section (Have you experienced a shock?) that will determine the enumeration of the rest of the section (if no shock, go to the next module);
  - When enumerating the shocks experienced, a question about each individual shock, or a general question, allowing the household to list all shocks;
  - Based on the suggested focus for this module, a filter question to ensure the HH understands they are responding with a list of shocks that have affected their economic status or their ability to acquire food;
  - Once the section on shock identification is complete, coping mechanisms used during the same period, ideally from a pre-coded list (i.e. the enumerator lists all the coping mechanisms used during the recall period).
- Rank the importance of a given shock for that household.
- Record the seasonality of the shock or coping mechanisms. When during the

previous recall period did the household experience this shock or use the coping mechanisms? (Note that this question has limited field experience but can have interesting results.)
- Link each specific shock (two to four main shocks) to a coping mechanism. However, feedback suggests that unless the problem is major – for example, an earthquake – households often struggle to link one unique coping strategy to one unique shock.
- Instead of 12 months, the recall period, if in an FSMS, may be back to the previous round of data collection. In an EFSA, it may be back to the period of the last main covariate shock.
- Relate specific shocks to their impact on assets (loss of, recovery of), usually in conjunction with linking specific shocks to coping strategies.
- Relate specific shocks to their impact on income (reduction of, return to the same level of).
- Relate specific shocks to the ability to feed the household.
- Relate specific shocks to current recovery status (not recovered, partially recovered, totally recovered).

Sources of inspiration

Questions on shocks and coping strategies can be found in the LSMS questionnaires. Such questions are quite similar to those included in the CFSVA module, even though they draw attention to the impact on household welfare, whereas the CFSVA module looks more specifically at the impact on food consumption.

Table 4.14 presents a typical set of questions and options for the Shocks and Coping Strategies module. However, note that analysing the number of indicators presented is difficult, as most of the data obtained from such a table are not significant due to the multiple combinations of answers that result in a small number of cases. Hence, it is suggested to group the indicators based on type. For example, "Reduced number of meals," "Reduced proportions of the meals," "Rely on less preferred, less expensive food," "Reduced expenditures on health and education" can be grouped as "adjustment strategies," while all borrowing can be combined as "borrowing strategies." Similarly, all selling of assets could be combined as "divestment strategy," and all instances of migration could be combined as "migration strategy." The analyst can create other categories based on the importance of a particular coping strategy in a particular country, and on the number of households that adopted the strategy.

Links with other modules
- This section should not be combined with the Coping Strategies Index (CSI), nor is it a substitute for a CSI (a CSI is done in addition to this and has its own specific methodology).
- To the qualitative data, if collected.

Tips
- If using the initial skipping rule, be sure the enumerators do not lead respondents on this question in order to skip the entire section.
- It is imperative that a relevant list of shocks and food and non-food coping strategies is created.
- When linking shocks and coping strategies, enumerator training and clarification of questions is important.

Table 4.14: Example of Shocks and Coping Strategies module

		1. Rank three shocks 1 = main 2 = second 3 = third	2. Did [shock] decrease your household's ability to produce or purchase food? 1 = Yes 2 = No 3 = Don't know
Has your household experienced any of these shocks that have made it difficult to obtain sufficient means of livelihood in the last 12 months? *If yes, please rank the shocks and report the three most serious.* *If no shock affected the household, go to question 9.3.*			
a	Drought/irregular rains	|__|	|__|
b	Regular floods	|__|	|__|
c	Flash floods	|__|	|__|
d	Landslides, erosion	|__|	|__|
e	Severely high level of crop pests and disease	|__|	|__|
f	Severely high level of livestock diseases	|__|	|__|
g	Lack or loss of employment	|__|	|__|
h	Unusually high level of human disease	|__|	|__|
i	Fire	|__|	|__|
j	High costs of agricultural inputs (seed, fertilizer, etc.)	|__|	|__|
l	Earthquake	|__|	|__|
m	Reduced income of a household member	|__|	|__|
n	Serious illness or accident of household member	|__|	|__|
o	Death of a working household member	|__|	|__|
p	Death of other household member	|__|	|__|
q	Theft of money/valuables	|__|	|__|
r	Theft of animals	|__|	|__|
s	Conflict	|__|	|__|
t	Other (Specify:_____)	|__|	|__|
3.	What did the household do to compensate for this loss of income and/or assets? (Report the three most important coping strategies. Choose from the strategy codes below).	|__|__| 1st |__|__| 2nd |__|__| 3rd	

Coping strategies codes:

01 Rely on less preferred, less expensive food
02 Borrowed food, helped by relatives
03 Purchased food on credit
04 Consumed seed stock held for next season
05 Reduced the proportions of the meals
06 Reduced number of meals per day
07 Skipped days without eating
08 Some HH members migrated
09 Sold durable household goods
10 Sent children to live with relatives
11 Reduced expenditures on health and education
12 Spent savings
13 Gathering
14 Sold or consumed livestock
15 Sold agricultural tools, seeds, or other inputs
16 Worked for food only
17 Sold crop before harvest
18 Rented out land
19 Sold land
20 Borrowed money
21 Other (Specify: _____)

Module Title: Coping Strategies Index

Main purpose of the module: The CSI is a relatively simple and easy-to-use indicator of household food security; it is straightforward and correlates well with more complex measures of food security. A series of questions about how households manage to cope with a shortfall in food for consumption results in a simple numeric score. In its simplest form, monitoring changes in the CSI score indicates whether a household's food security status is declining or improving.

Limitations of the module

See the limitations discussed under the Shocks and Coping Strategies module.

Creation of the Coping Strategies Index module

The question to ask is **"What do you do when you do not have enough food, and do not have enough money to buy food?"** The answers to this question are the basis for the CSI module.

- The minimum structure of this module should be a list of strategies used to cope with the food shortage or when households do not have enough money to buy food.
- The recall period for CSI is recommended to be the past seven days.

One category should be "daily" or "all the time," and one category should be "never." The intermediate categories can be changed around according to conditions and the amount of detail required. In general, at least five relative frequency categories are recommended, as shown in Table 4.15.

When using the CSI, the question at the top should be repeated for each of the strategies on the list, and the appropriate relative frequency box should be ticked.

The best way to assess the frequency of coping strategies is not to count the number of times a household has used them, but to ask a household respondent for a rough indication of the relative frequency of their use over the previous month. Precise recall is often difficult over a long period of time, and asking for the relative frequency provides adequate information. There are various ways a relative frequency count can work – this one asks roughly what proportion of the days of a week people had to rely on various strategies.

Modification to the module

- Although a generic list of strategies is presented in Table 4.15, list only those strategies applicable to the area. There is no point in asking about the strategies not adopted in the area (for instance, in a non-agricultural population, we do not need to ask about "consuming the seed stock").

Table 4.15: Example of questions for constructing a CSI

In the past 30 days, if there have been times when you did not have enough food or money to buy food, how often has your household had to:	Never	Seldom (< 1 day a week)	Sometimes (1–2 days a week)	Often (3 or more days a week)	Daily
a. Reduce number of meals eaten per day?	1	2	3	4	5
c. Borrow food or rely on help from friends or relatives?	1	2	3	4	5
d. Rely on less expensive or less preferred foods?	1	2	3	4	5
e. Purchase/borrow food on credit?	1	2	3	4	5
f. Gather unusual types or amounts of wild food/hunt?	1	2	3	4	5
g. Have household members eat at relatives' or neighbours'?	1	2	3	4	5
h. Reduce adult consumption so children can eat?	1	2	3	4	5
i. Rely on casual labour for food?	1	2	3	4	5
j. Feed working members of HH at the expense of non-working members?	1	2	3	4	5
k. Go entire day without eating	1	2	3	4	5
l. Consume seed stock to be saved for next season	1	2	3	4	5

The context-specific CSI "has been criticized for being relatively unhelpful in comparative analysis. However, fieldworkers have noted that several of the individual behaviours that the CSI measured recur across different contexts"[60]. Recognizing this, a reduced CSI was developed to compare food security across different contexts. It is based on the same short list of (5) coping strategies (see figure below) and the same severity weights.

The reduced CSI is less valuable in identifying the most vulnerable HHs in a given location, but it is very useful for comparison across countries as it focuses on the same set of behaviours. Furthermore, "extensive research demonstrated that the 'reduced' CSI reflects food insecurity nearly as well as the 'full' or context-specific CSI".[61]

The figure below describes how to calculate a reduced household CSI score:

In the past 7 days, if there have been times when you did not have enough food or money to buy food, how often has your household had to:	**Raw Score**	**Universal Severity Weight**	**Weighted Score = Frequency X weight**
Relative Frequency Score			
a. Rely on less preferred and less expensive foods?	5	1	5
b. Borrow food, or rely on help from a friend or relative?	2	2	4
c. Limit portion size at mealtimes?	7	1	7
d. Restrict consumption by adults in order for small children to eat?	2	3	6
e. Reduce number of meals eaten in a day?	5	1	5
TOTAL HOUSEHOLD SCORE – Reduced CSI	Sum down the totals for each individual strategy		**27**

In order to conduct the analysis of the CSI, however, you need a few more pieces of information. The first is a way to "weight" the individual strategies or behaviours.

The CSI tool relies on counting coping strategies that are not equal in severity. Different strategies are therefore "weighted" – multiplied by a weight that reflects their severity before being added together. The simplest procedure for doing this is to group the strategies according to similar levels of severity and assign a weight to each group. The severity of coping strategies is, to some extent, a matter of perception.

Focus group discussions with different community groups are needed to determine the severity of the coping mechanisms. The first step is to try to group the strategies into categories of roughly the same level of severity. Since this task is carried out with different groups, it is useful to impose some structure from the outset. For example, one could divide them into four different categories: very severe, severe, moderate, and not severe.

- It is always easiest to determine the most severe coping strategies, so ask the group to select the most severe and least severe individual strategies first.
- Then ask if there are other individual strategies that are more or less the equivalent of these two, in terms of how severe they are perceived to be. Once the two extremes are established, it is easier to group the remaining behaviours into intermediate categories.

60. Source: "Coping Strategy Index: Field Methods Manual" II edition (2008)
61. ibid

- This must be done with enough groups to represent the diversity of the location or culture, to ensure that a reasonable consensus has emerged.
- Incorrectly weighting individual strategies will result in errors in the analysis.

Table 4.16: Example of coping strategies grouped and ranked by focus group

Strategy	FG1	FG2	FG3	FG4	FG5	FG6	FG7	FG8	Avg	Rank
a. Limit portion size	1	1	1	1	1	1	1	1	1	1
b. Reduce number of meals	1	2	1	1	3	1	1	1	1.4	1
c. Borrow food	2	2	3	3	2	2	2	3	2.4	2
d. Rely on less preferred/expensive foods	1	1	1	1	1	1	1	1	1	1
e. Purchase/borrow food	2	2	2	3	2	2	2	2	2.1	2
f. Gather unusual types	5	5	4	4	3	5	5	5	4.5	5
g. Eat at relatives' or neighbours'?	2	n/a	n/a	2	2	3	2	n/a	2.2	2
h. Reduce adult consumption	2	3	2	2	3	3	2	2	2.4	2
i. Rely on casual labour	1	1	2	2	1	-	2	2	1.6	2
j. Feed working members	3	3	3	4	3	3	5	3	3.4	3
k. Go entire day without eating	5	4	5	5	5	4	5	5	4.8	5
l. Consume seed stalk	4	3	4	3	3	3	4	4	3.5	4

Source of inspiration

- The CSI is consistent with the CARE/WFP methodology.

4.2.4.3 Non-core modules with non-changeable questions

Module Title: Maternal Health and Nutrition

Main purpose of the module: The aim of this module is to gather data about the health and nutrition status of mothers with children under 5 years of age. The information can be used to understand the relationship between malnutrition, diseases, and consumption.

Limitation of the module

- Health status is usually self-reported by women and not clinically demonstrated.

Creation of the Maternal Health and Nutrition module

Weight, height, and mid-upper arm circumference (MUAC) are collected to calculate the woman's nutrition status.

Measurements of mother		
11.22	**Mother's height** (in centimetres)	|__|__|__|.|__|cm
11.23	**Mother's weight** (in kilograms)	|__|__|__|.|__|kg
11.24	**Mother's MUAC** (in centimetres)	|__|__|.|__|cm

Health is commonly assessed through questions about illness or diseases (fever, diarrhoea). However, many additional questions may be included.

Modification to the module

- Additional questions about a mother's health can be added: Is she currently pregnant or breastfeeding? The number of pregnancies, miscarriages, stillbirths? The number of living children? The age of the first delivery? Has she received iron-folate tablets, antenatal care, vitamin A capsules? What are her health and hygiene practices (sleeping under a mosquito net, boiling or treating drinking water, washing hands before preparing meals or after going to the toilet)? Her level of education, occupation, and control over food and income could be included as potentially relevant for the analysis of the results.

Links with other modules

- Child health and nutrition
- Water and sanitation
- Food consumption
- Community-level infrastructure (specifically, health facilities)

Source of inspiration

- MICS and DHS surveys have more extensive questionnaires for women, including information on child mortality, maternal and newborn health, and marriage. Likewise, the CFSVAs, MICS, and DHS collect anthropometric data on women.

Module Title: Child Health and Nutrition

Main purpose of the module: The aim of the module is to gather data about the nutrition status and health conditions of children under 5 years old (usually 6 to 59 months).

Limitations of the module

- Appropriate sampling approaches and sample sizes are required to calculate prevalence rates with a sufficient degree of precision.
- If that is not feasible, data can be collected and used to investigate relationships between nutritional outcome and other food security indicators, but not to provide prevalence rates.
- Collecting height and weight data requires some effort. Data are collected using special equipment that is bulky and troublesome to carry around to each household. Frequently, an individual appropriately trained in anthropometry must be added to each survey field team.

Creation of the Child Health and Nutrition module

The key data to assess child nutrition status are: sex, age, weight and height/length. Additionally, MUAC is often collected. Basic health information is related to illness, particularly to diarrhoea and fever. Table 4.17 offers an example of a simple Child Health and Nutrition module.

Table 4.17: Example of Child Health and Nutrition module

ASK SELECTED RESPONDENT IF THERE ARE CHILDREN OF 6–59 MONTHS OF AGE IN THE HOUSEHOLD. IF NOT, TERMINATE INTERVIEW.
Read aloud: Now I would like to ask you some questions about the children in this household (Continue the interview with the same woman)
We would like you to come with all the children aged 6 to 59 months from your household. We would like to measure and weigh them as part of our assessment.
It is very important that children are measured, so be persuasive. Assist women in transportation, if need be.

Starting with the youngest child, and focusing on one child at a time, enter each child's first name and ask for the following information:

1.a	1.b	2.	3.	4.	5.	6.	7.	8.	9.	10.
First name	Mother's/ caretaker's ID no. (link with mother's section, if collected) 8 = missing at interview 9 = dead	If available, date of birth from the medical card If no → 10.3 If yes → enter, then → 10.4 Use format dd/mm/yy	Child's age in months	Child's sex? 1= Male 2 = Female	Are you the mother/care-taker of [Name] 1 = yes 2 = no If no → 10.8	Has [Name] been ill in the last 2 weeks? 1 = yes 2 = no → 10.8 3 = do not know → 10.8	Has [NAME] been ill with diarrhoea and/or fever at any time in the past 2 weeks? (Diarrhoea: perceived by mother as 3 or more loose stools per day for 3 days or one large watery stool or blood in stool) 1 = yes 2 = no 3 = do not know Diarrhoea Fever	Child's height/ length (in centimetres, with 1 decimal place)	Child's weight Enter weight in kilograms, with one decimal place	Child's MUAC (in centimetres, with 1 decimal place)
	__	__/__/__	__	__	__	__	__ __	__.__	__.__	__.__

Modification to the module

- Additional questions about the child's health can be added: size at birth, breastfeeding and weaning history, vaccination, de-worming or other treatment. Also useful would be questions on other diseases (e.g. malaria, measles, acute respiratory infections) and on the child's feeding patterns.

Links with other modules

- Mother health and nutrition
- Water and sanitation
- Food consumption
- Community-level infrastructure (specifically, health facilities)
- Sources of inspiration
- MICS and DHS surveys have more extensive questionnaires for children. Similar to the CFSVAs, MICS and DHS collect anthropometric data on children and information on health status. LSMS questionnaires sometimes include a children's module for anthropometric data.

4.2.4.4 Non-core modules with changeable/flexible questions

Gender-sensitive survey design and implementation

Study preparation

An in-depth literature review can be used to identify factors that shape gender relations, such as cultural beliefs, values and practices, religion, education, politics,

legislation, economic situation and demographic factors. Generating this type of overview prior to primary data collection provides a context for tailoring generic data collection tools to ensure that they are gender-sensitive and appropriate for a particular setting.

Selection, composition, and training of survey teams

Although members of field teams do not need to have a technical background in gender analysis per se, it is crucial that enumerators are sensitized to the importance and rationale behind collecting sex-disaggregated data and phrasing questions in a way that allows for an analysis of the relationship between gender, food security, and vulnerability. This is even more important for facilitators applying qualitative tools such as focus group discussions and participatory rural appraisal techniques. A balanced mixture of male and female enumerators will minimize the extent to which bias is introduced due to enumerator gender. Where group discussions are to be held separately for men and women, same-sex discussion facilitators are likely to contribute to a more relaxed and open discussion.

Study design (household surveys)

In each questionnaire, sex, age, and relationship to the household head of the main respondent should be indicated to determine possible biases introduced during the data collection process. This will also assist with the identification of different perceptions of men, women, and age groups during analysis.

To the extent possible, all questions concerning food security and vulnerability included in household surveys should be designed in such a way as to differentiate between the experiences of women and men (girls and boys). Please refer to Box 4.6 for examples of key questions that should be disaggregated by gender. Gender-disaggregated data provides valuable information about those intra-household differences that can be masked by surveys that treat households as a single, homogenous unit. Quantitative indicators produced by household survey data can be used to measure the degree of gender inequalities related to food security and vulnerability.

Community discussions or interviews

The qualitative data generated through discussions or interviews with community members provides key insights for understanding the underlying causes and reasons for inequalities identified during household surveys, and allows for further elaboration of the causal mechanisms suggested by quantitative data.

Interviews with key informants (i.e. local authorities)

It is important to include knowledgeable women in the list of persons to be used as key informants. Women's organizations or women's affairs offices often provide suitable candidates. Discrepancies between authorities' perceptions and household- and community-level realities enable an assessment of whether key decision-makers are aware of gender-related differences and inequalities.

Timing of data collection

Appropriate timing is crucial for ensuring that women and men are able to participate in all data collection exercises. Although communities are busy throughout the year, there may be periods when their workload is slightly less burdensome. Similarly, the

availability of community and household members is influenced by the daily pattern of agricultural work, and the income-generating and household activities of men and women. For example, women may not be able to attend meetings during evening hours due to domestic responsibilities.

Box 4.6: Key questions for use in conducting gender analysis (WFP 2006)

Household Roles/Social and Cultural Constraints

- What are the different needs, roles, and interests of women and men?
- What are the power dynamics between women and men?
- Which decisions are made by men and which by women?
- What are the social and cultural constraints and opportunities of women and men?
- What are the relations between women and men in society, the community, and the household?
- What different coping mechanisms are available to women and men to lessen the risk of food insecurity for their families?
- How do access to and control of resources, information, and services affect participation by women and men in the programme/project?
- How do gender roles (e.g. workload, time, mobility) influence the ability of women and men to participate in the programme/project?

Food and Livelihoods

Who manages food within the household?

- How is food distributed within the household?
- Who cultivates land and grows food?
- Who is the family's main income earner?
- What are the income-generating opportunities and needs of men and women?
- Where is it convenient for women and/or men to collect food assistance?
- Who collects food assistance?

Health Risks and Accessibility to Health Services

- What are the health risks for women and men? How and why are they different?
- What barriers (e.g., self-confidence, mobility, financial resources, role in decision-making) do women and men face in accessing health services and health information?
- Where do women and men go for health services and information?
- Which communication channels are most appropriate for women and men?
- Can women and men discuss their health problems/issues among themselves?
- Is this culturally accepted?
- Where can women and men learn more about how to address their health concerns?
- What social networks exist in the community for men and for women?
- Can these networks help address health concerns?

HIV-Affected Households

- For HIV-affected households, what are the different coping mechanisms of women and men? Of girls and boys?
- For HIV-affected households, what is the impact on girls' and boys' school attendance? Are girls withdrawn from school more often than boys?
- What are women's and men's responsibilities related to caring for PLHIV?

Collecting data on gender issues

As a general rule, gender-sensitive data is to be collected in each module of a household survey. In the context of a CFSVA, the gender-sensitive data usually incorporated (or that can be easily incorporated) into a household survey includes:

- **Demography:** sex and age of the head of the household; household composition;
- **Migration:** circumstances of migration (reasons, remittances, gender of the migrated members) in order to assess impact of migration on gender prevalence (at the household and community levels) and on food security;
- **Education:** primary and secondary school attendance of girls and boys, and literacy skills of the head and his/her spouse, informal training of men and women;
- **Income sources:** differentiated participation of household members in income-generating activities;
- **Food consumption:** intra-household distributions and sequence of family members eating food; and
- **Health and nutrition:** prevalence of malnutrition among women, prevalence of child malnutrition by gender, data on breastfeeding and reproductive health, and awareness of HIV/AIDS prevention and transmission.

Key questions in the household questionnaire for HIV/AIDS

Literature on HIV/AIDS identifies some key attributes of chronically ill or deceased adults that are crucial to studying the impact of and responses to HIV/AIDS. These attributes include age, gender, relationship to the household head, educational level, active role of the individual in the household, and decreased capability to work. Ideally, a survey on the impact of HIV/AIDS should collect information on all these attributes. Within the context of food security assessments, the minimum set of attributes to consider includes:

- age;
- relationship to the head of the household; and
- decreased capability to work.

The way we capture the presence and key attributes of chronically ill or deceased household members depends on how demographic data are collected during the household survey.

Option 1: Data are collected through a roster

If data are collected through a roster, ID, name, age, gender, and relationship to the household head are typically collected. Table 4.18 shows how a roster can be adapted to capture information related to chronically ill members. Yellow highlighted sections help identify chronically ill adult members and some key attributes.

Table 4.18: Demographic data collected through a roster

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
ID	**Name**	**Age** (in years, if <1yr. old, write 0)	**Gender**	**Relationship to the HH head**	**Has s/he been not fully functional for at least 3 months over the past 12 months?**	If yes, **with which kind of illness?**	**Is s/he engaged in paid work (cash/in-kind)?** If ill, consider the period before illness.	If chronically ill, **over the past 12 months, has s/he been able to work as before?**
01 02 03 04 05 ... ... N	____ ____ ____ ____ ____ ____ ____ ____	|__||__| |__||__| |__||__| |__||__| |__||__| |__||__| |__||__| |__||__|	|__| |__| |__| |__| |__| |__| |__| |__|	|__| |__| |__| |__| |__| |__| |__| |__|	|__| |__| |__| |__| |__| |__| |__| |__|	|__| |__| |__| |__| |__| |__| |__| |__|	|__| |__| |__| |__| |__| |__| |__| |__|	|__| |__| |__| |__| |__| |__| |__| |__|
		If 98 or more, write 98 99 = NK	0 = M 1 = F	1 = head 2 = spouse 3 = son/daughter 4 = father/mother 5 = brother/sister 6 = grandparent 7 = uncle/aunt/ cousin 8 = niece/nephew/ grandchild 9 = adopted/foster child 10 = stepchild 11 = no relation	0=no 1=yes	0=mentally/ physically disabled 1= chronic illness	0=no 1=yes	1=yes, able to work the same number of hours/days 2= no, working for fewer hours/days 3=completely unable to work

NK = not known

Data on deceased members require nesting a separate table within the questionnaire. Table 4.19 suggests how questions on recent deaths can be formulated. They are very similar to the questions on chronically ill members.

Table 4.19: Collecting data on deceased household members

Q1	Did any adult (ages 18–59) household member die during the 12 months before the survey after being sick for at least 3 months over the past 12?				|__| 0=no (skip the whole section) 1=yes
For each of the adult (ages 18–59) household members who died after being sick for at least 3 months over the past 12, report:					
	Q2 **Cause of death**	**Q3** **Gender**	**Q4** **Relationship to the HH head**	**Q5** **Was s/he engaged in paid work (cash/in kind)?** consider period before illness	**Q6** **In the period s/he was sick, was s/he able to work as before?**
1 2 ... N	|__| |__| |__| |__|	|__| |__| |__| |__|	|__| |__| |__| |__|	|__| |__| |__| |__|	|__| |__| |__| |__|
	1=after chronic illness 2=after a period of physical disability 3=old age 4=problems caused by pregnancy 9=other (Spec. :_________)	0=M 1=F	1=head 2=spouse 3=other member	0=no 1=yes	1=yes, able to work the same number of hours/days 2=no, working for fewer hours/days 3=completely unable to work

Option 2: Data are not collected through a roster

If the household questionnaire does not include a roster, questions on deceased members are the same as in Option 1; questions on chronically ill members need to be asked in a different format.

Table 4.20: Demographic data not collected through a roster

Q1	Has any adult (ages 18–59) household member been not fully functional for at least 3 months over the past 12 months?	|__| 0=no (skip the whole section) 1=yes

For each of the adult (ages 18–59) household members who have not been fully functional, report the following:

	Q2 **Kind of illness**	**Q3** **Gender**	**Q4** **Relationship to the HH head**	**Q5** **Was engaged in paid work (cash/in kind)?** Consider period before illness	**Q6** **Over the past 12 months has s/he been able to work as before?**
1 2 … N	|__| |__| |__| |__|	|__| |__| |__| |__|	|__| |__| |__| |__|	|__| |__| |__| |__|	|__| |__| |__| |__|
	0=mentally/ physically disabled 1=chronic illness	0=M 1=F	1=head 2=spouse 3=other	0=no 1=yes	1=yes, able to work the same amount of hours/days 2= no, working for fewer hours/days 3=completely unable to work

4.2.5 Ensuring data quality

The manager of the CFSVA should ensure that the highest quality data are collected. Data quality is influenced by many factors, such as whether or not PDAs or paper questionnaires are used, the selection of previously experienced enumerators (knowledge of local languages and the language of the supervisors, country, food security, data collection), the quality of the training, how motivated staff are to collect accurate data (this can be seen during the test), and how well the data collection process is supervised. If PDAs are used, enumerators should possess basic computer knowledge.

4.2.5.1 Training of the enumerators

Quality training is essential to ensure that enumerators are able to ask questions in a non-leading way, probe during the interviews, and triangulate the information collected.

Although there are several ways to go about enumerator training, there are some rules of thumb for CFSVAs. A typical training schedule could include sections on:

- Administrative issues;
- Overview of survey purpose and objectives;
- The role of interviewers and interviewer comportment, standard operating procedure (SOP);
- Overview of questionnaire and discussion of individual items;
- Mock interview demonstrations;

- Role playing;
- Field exercise; and
- After action and review.

For the main survey instruments (household and key informant, not including child anthropometry), training usually takes about five to eight days.

- three to four days of in-class training;
- one to two days of field testing; and
- one to two days to review issues after the field testing.

- If the survey includes child anthropometry data collection, an additional two to five days will be needed for the enumerators responsible for the child nutrition data collection, depending on the level of previous experience, including field testing.
- If PDAs are to be used, an additional one to three days of training will be necessary, depending on the level of computer experience the enumerators have. It is also recommended to have an ad hoc trainer for the PDA.
- If qualitative methods (focus groups) are part of the survey, an additional two to eight days will be needed, depending on the skill level of the enumerators.
- Team leaders will need a separate one to two days of training on issues such as key informant questionnaires (if they are the ones responsible), selecting households, and reviewing their responsibilities.
- It is best to have two or more trainers, preferably including a national staff member.
- Try to have the questionnaires as finalized as possible before the training begins.
- In training, try to mix formal presentations with discussions and group work, practice sessions, and other examples to keep the attention of the participants.
- The trainers are responsible for modifying the questions in the questionnaires and/or the PDA file accordingly with the recommendations coming from the field test.
- If PDAs are used, the trainers, in collaboration with the supervisors, are responsible for the logistics (charging the batteries and updating the latest version of the questionnaire file on the PDAs).

4.2.5.2 Field testing

The field testing is a crucial, and mandatory, step in ensuring data quality. It involves checking the questionnaire, raising problems/issues that arise with the questions, and answering the doubts the enumerators can face. Once the field test is done, it is essential to modify the questionnaire in accordance with the observations and discussions made by and with the enumerators.

The rules of thumb for field-testing during a CFVSA

- Generally, this is done over one to two days. Often there are two phases: the second field test uses the final version of the questionnaire and captures the latest mistakes and difficulties encountered by enumerators.
- Each enumerator should be responsible for a minimum of two household interviews, preferably more.
- It would be useful to send the enumerators to do the interviews in pairs, with one asking the questions and the other listening. After the first interview they can exchange roles. This process helps the enumerators better understand how to ask the questions and allows them to give their opinion.

- The team leaders should work with their assigned teams. They should also be field-testing their supervisory responsibilities.
- Pick a site for field-testing that is typical of real field conditions yet relatively easy to reach. The field test should be as similar to the real field conditions as possible.
- It is often preferred not to inform the community and informants they are only participating in a "test." Even if they are not included in the sample, they are filling an important role as the field test site, which will improve the quality of the survey, much as the participation of any community and informant selected in the sample will improve the quality of the survey.
- If child anthropometry data collection is expected, the field test is a good opportunity to check if the enumerators are able to perform the measurements under field conditions.
- If PDAs are used for data collection, it is advisable that the PDA trainers join the groups. This way they will immediately address issues that might come out during the test. This is particularly important if GPS units are used.

4.2.5.3 Field supervision

A daily quality control of the data collected is highly recommended whether paper questionnaires or PDAs are used. Having a replacement for any enumerator (or supervisor) who may become sick, who leaves for personal reason, or who is not adapted to the type of work required is also essential to the quality of the data. A supervisor oversees all or part of the enumerator teams; a team leader is the head of one team of enumerators. Supervisors should conduct spot visits of the data collection teams in the field in order to check on the work achieved, especially during the beginning of the survey. They should also revisit a random sample of households to ensure quality.

Team leader responsibilities

General

The team leader is in charge of managing all data collection, starting from the selection of the households, to the assignation of households to each enumerator, and the checking of each questionnaire filled out.

The team leader should support the enumerators during data collection and clarify and resolve issues or problems raised during data collection. He/she should also be in charge of the time, planning and managing data collection based on the number of hours the team will stay in the field. He checks the enumerators' work by sitting in on interviews, reading completed questionnaires for misunderstandings and recording errors, liaising and troubleshooting with respondents and local community leaders, and covering for enumerators in emergencies (e.g. arranging interviews if an enumerator falls sick).

Paper questionnaire

The supervisor should be able to check each questionnaire in the field before leaving the village, so any mistakes or missing information can be corrected/filled in by the enumerators. By marking any errors with a red pen, the supervisor makes them more visible and reminds enumerators not to repeat those mistakes.

The supervisor should also check that the numbering of the questionnaire questions is correct; he/she is also in charge of collecting all the questionnaires at the end of the day and storing them for the data entry process.

PDA
The supervisor should check the questionnaires for data quality. The PDA software allows supervisors to load the questionnaires saved on each SD (memory) card so that they can review entered data. Supervisors are also responsible for assigning the PDAs to the enumerators.

Data is automatically copied in two locations by the survey software: SD card and PDA internal memory. At the end of each day the files should be copied to a PC, if available, as a further backup.

Enumerator responsibilities

General
The enumerators are the first responsible for the quality of the data collected. A job description for an ideal enumerator would include: communication skills, good knowledge of the international language of the country (English or French, or Portuguese or Spanish) as well as the local language(s), a perceptive intelligence, inexhaustible patience, unfailing dependability, outstanding people skills, and a willingness to work long hours.

Documents
Each enumerator needs to bring enough copies of the questionnaire for the whole day, plus some spare copies. If pens are used to fill in the questionnaires, they should bring extras. If the enumerators prefer pencils, they need to bring erasers and a pencil sharpener.

Once the data collection is concluded, the questionnaire should be kept in a protective plastic folder. Each enumerator is responsible for bringing the questionnaires to the supervisor, checking for mistakes with him/her, and making the necessary changes.

Legible handwriting is important. When the enumerator checks an answer or writes it out, he/she should bear in mind that the data will be entered by a different person, who will need to be able to read the handwriting.

PDA management
If PDAs are used, the team leaders should be trained in how to manage them in order to provide support to enumerators during data collection. Necessary knowledge includes: how to maximize battery life, how to reset the PDAs, and how to reinstall the survey software, if needed.

Logistics for data collection

General
Implementing a successful assessment requires logistics planning and preparation. Logistics is an important part of the survey, and in some countries it can be a cumbersome exercise, and so should be considered early on in the process. It is linked

to selection of field sites and advance notification of sample communities or sites, coordination of transport and communications operations, and distribution and collection of data collection instruments.

Paper questionnaire
To ensure proper data collection, the teams should arrive in the field with the adequate equipment, including:
- Enough copies of the questionnaire
- Pens
- Pencils (rubbers and pencil sharpener)
- Seasonal calendar

If the CFSVA includes anthropometric measurements, the enumerators taking the measurements are responsible for bringing along the necessary equipment:
- Scale (UNICEF standards)
- Height board (children and women)
- MUAC tape

PDA transport
It is important to inventory every item related to the PDA before and after travelling. The list of necessary items generally includes:
- PDAs and manuals
- Chargers (one per PDA)
- Batteries (one per PDA)
- Stylus
- Storage cards
- GPS units, if needed
- Car chargers, if needed (in zones with no power supply)
- SD card reader for installing the software (usually one is sufficient)

PDAs are delicate hardware. They should be stored in a durable bag for transportation, to avoid their getting crushed by other cargo. PDAs usually come with a cover for protection, which should be used to avoid damage to the PDA.

After each day of use, the PDA's battery should be fully charged. If electricity is not available, team leaders should be provided with car chargers. Some PDA models can be charged via a USB cable attached to a laptop or a desktop PC. If GPS units are used, these should be charged as well. Charging PDAs and GPS devices requires about two hours.

When using the car charger, it is important to follow these steps:
1. Before connecting the converter, start the car's engine.
2. Verify that the switch at the rear of the converter is off.
3. Plug the converter into the cigarette lighter.
4. Make sure the switch on the multiple connectors is off.
5. Connect the multiples connectors to the converter.
6. Switch the converter on.
7. Turn on the multiple connectors' switch.
8. Start connecting the processor for PDA and GPS.

9. The unit is charged when the light is green.
10. Extinguish the multiple connectors and then plug in the converters.
11. Unplug the converter from the cigarette lighter.
12. Run the engine for another 20 to 30 minutes before shutting it down.

Managing problems with PDAs

Battery problems
With normal continuous usage, the battery lasts about 5 hours. In order to maximize the battery life:
- Charge the PDA as often as possible.
- Keep the screen brightness low.
- Turn off the volume.
- Switch off the PDA when not used.
- Enable the automatic "power off" function in the PDA's battery settings.

Resetting
If the PDA freezes, it is possible to do a "soft reset" to restart it (Note: you will lose the data that is currently being entered.) In most models, there is a button on the back of the PDA that you can push using the stylus. If this button cannot be found, consult the manual that comes with the PDA.

Speed/memory
To increase performance of the application, quit unused applications using the PDAs memory manager tool.

4.2.6 Key references: Survey

- Devereux and Hoddinott. 1992. *Fieldwork in developing countries*.
- CARE, Second Edition. 2008. *The Coping Strategies Index: Field Methods Manual.*
- WFP Nutrition Service. 2005. *Measuring and Interpreting Malnutrition and Mortality.*

4.3 HOUSEHOLD DATA ENTRY

The analysis of household survey data is undertaken in Microsoft Excel, SPSS, or STATA. A key step in the process is to convert responses collected during the interview into a format that can be easily manipulated by the analyst. The two most popular choices currently used by WFP are direct data entry through PDAs and manual entry of responses into a data entry programme. This chapter focuses solely on manual data entry by data entry operators using desktop or laptop computers. Regardless of the method employed to digitize the responses, a series of standardized steps must be followed to ensure the quality of data. The process may vary depending on the context, availability of resources, and the circumstances.

Ideally the management of quantitative survey data, or closed question/semi-open question qualitative data survey, will take the following steps:
1. Prior to the study, review the questionnaire and ensure that the format is adapted for ease of data entry.

2. Set up the data management entry process before the paper questionnaire is used.
3. Development of a data entry method, usually using a data entry programme, if large amounts of data are to be organized into a database.
4. Conduct data entry and check the accuracy.

Different data collection methodologies require the use of different questionnaires. Likewise, the types of data that need to be captured will vary by type of survey. In close-ended (quantitative) questionnaires, coded categorical responses, yes/no answers, and figures are entered into a database. For qualitative studies, answers in the form of narratives need to be coded and then entered. For the purpose of a CFSVA, regardless of the data entry application employed, the output or captured data must be easily transferable into SPSS, STATA, or Microsoft Excel for analysis.

4.3.1 Review of the questionnaire

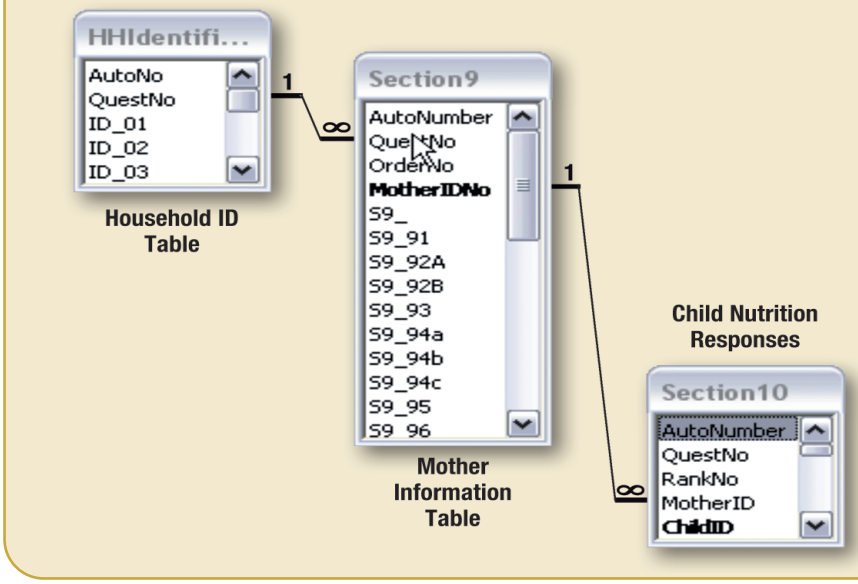
During questionnaire design, the person responsible for designing the data entry tool (who should be familiar with the software used) must review the format of the household and village questionnaire to ensure that all questions are properly numbered; that the layout of the questionnaire allows for easy construction of the relationships between the tables, and that unique ID numbers are logically constructed based on an agreed-upon coding system.

Each question and table should be clearly and logically numbered. In addition, the unique ID code assigned to each household should be clearly labelled.

4.3.1.1 Household and household member relationships

An important step in the design of the household questionnaire is creating relationship tables between the differing hierarchies within the households.

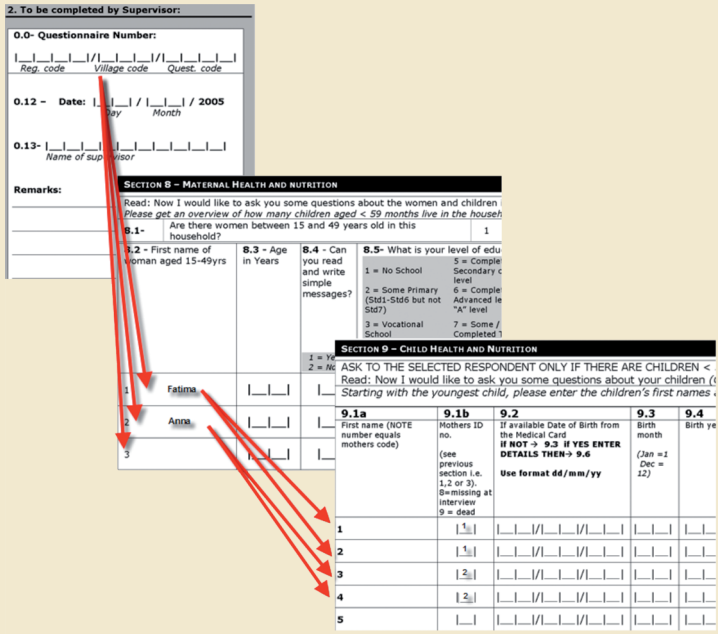
Figure 4.4: Household and member relationship tables



This is particularly important if the study is attempting to record the responses of the household, household members, mothers in the household, and information on each child. In terms of data relationships, the household is the highest unit; the household members and mothers have "one-to-many" relationships to the household (e.g. one unique household can have many members, each with his/her own characteristics), and the child has a "one-to-many" relationship to the mother. For example, in a single household (one unique HH ID), there are two mothers who have responses recorded in the survey (both have the same unique HH ID, but each has her own unique mother ID). Each of these mothers has two children under 5 whose age, weight, and height are measured (each child has the same unique HH ID, each set of siblings has the same mother ID, and each child has his/her own unique child ID). The household-to-mother and household-to-child relationship is one-to-many, the mother-to-child relationship is one-to-many, and the reverse relationships are many-to-one. Figure 4.4 is an example of the relationships from the East Timor CFSVA.

In order to facilitate the one-to-many relationships in a data entry application, it is much easier if the questionnaire uses a horizontal table, where each member of the sub-group (e.g. the children) is recorded on a horizontal line. The enumerator can add as many lines as required in the questionnaire based on the number of children in the household concerned; the programme can also insert the same number of lines using one-to-many relationship tables. The example from the Tanzania Household CFSVA questionnaire in Figure 4.5 shows the relationship.

Figure 4.5: One-to-many Relationships



4.3.2 Paper questionnaire management before data entry

Organizing filled-in questionnaires in a systematic fashion will prevent questionnaire loss. Good questionnaire management is a prerequisite for ensuring timely data entry. The steps necessary to develop an effective questionnaire management system at the data processing office are discussed here.

1. As envelopes/boxes of questionnaires return from the field, organize them into groups. The envelopes can then be further organized into larger groupings, usually by province or district. For example, all questionnaires from the same village/cluster can be put together in one envelope, with the name of the village, district, province, etc. written clearly on the envelope. Then all the envelopes from the same district can be put together in one box, with the names and codes of the province and district written on the box. Be sure that the system of organization uses information that is also included on the questionnaire and entered into the database, so that if a physical questionnaire needs to be located, the information leading to the correct envelope will be contained in the database.

Box 4.7: Correcting errors on paper

- When correcting errors or making changes to completed questionnaires, use a red pen to make it clear what is original data and what has been added or changed. Give red pens only to those responsible for making changes/corrections.
- Always cross out bad answers with one slash mark. Do not scratch out answers, or use liquid paper or erasers. This way, should a correction be made in error, the original answer will still be legible.
- Never fill in missing answers unless it is 100 percent clear what the answer should be. It is better to have missing data than incorrect data!

2. Each questionnaire should have a unique ID, i.e. a number that identifies a questionnaire from all other questionnaires in the survey. This number can be assigned at any time, but is best done before the teams are in the field. It is recommended that a questionnaire number be used. The questionnaire number is a sequential number that can be constructed with each team starting off with a number and then continuing in continuous order (e.g. 1001,1002, . . . 1999)

Example of Questionnaire Identification Number

QUESTIONNAIRE ID:
|_|_|/|_|/|_|/|_|_|/|_|_|/|_|_|
Prov. Dist. Terr Sec. Group. Men

3. As questionnaires are entered into the database, they should be filed in sequence according to their unique ID numbers.
4. Sometimes it is helpful to have a separate group of people (not the enumerator or data entry operators) responsible for carefully reviewing every

paper questionnaire, looking for problems, making corrections where possible, and ensuring that all answers and unique IDs are clearly recorded. The decision to add this step depends on the type of questionnaires and questions being used, the number of errors present, the recoding needed, and the experience and understanding of the team leaders and data entry operators. If the team leaders and data entry operators have the experience to identify and correct most of the mistakes, then this separate step might not add much value. Correction of mistakes should be done in a clear manner, as sometimes the corrections themselves can lead to confusion for the data entry operators.

When data is gathered at several levels in one survey, a decision must be made regarding how to organize the different levels of information. One option is to keep all the questionnaires together based only on the geographic or other filing system. Another is to separate the questionnaires based on level of data collection (household, focus group, key informant, etc.), and file them separately.

4.3.3 Development of a data entry application or programme

4.3.3.1 Data entry applications

Principles

Developing data entry applications can be complex and time-consuming. However, time spent developing an application will mean fewer errors at the data entry stage. Although each application requires some customization, the key principles in any data entry application are similar:

1. Maintain household records and the relationships between the sub-household units (household members, household mothers, and children).
2. Create data entry masks (forms), which allow the data entry person to enter records in a standardized and intuitive way to minimize key strokes and mouse clicks.
3. Use structured entries that unambiguously link every household to its administrative units and limit the ability of data entry persons to input improbable values and categories, and also use embedded logic commands (such as filters) that control for simple data entry errors.
4. Export entered data into statistical software to ensure minimal effort and no loss of integrity.

Keeping these principles in mind, this section guides the design of data entry templates that adhere to these four principles while allowing for the specific requirements of each study. This guideline recommends Microsoft Access, which has been used in this example, as the platform for data entry. However, many other software packages are available and are commonly used.

4.3.3.2 Software for data entry and management

A number of software packages are available to facilitate data entry into a computer system. Although this guideline uses Microsoft Access to show how to develop a data entry template, data mask, and how to enter data, there are other applications available which can be used for data entry. Brief descriptions of commonly used applications are provided below.

Microsoft Access

Access is a Microsoft database application that allows data to be stored in related tables based on unique IDs. Data entry in MS Access can be undertaken in designed forms. The forms can be customized with embedded Visual Basic for Application (VBA) code to allow filters and logic commands to control for clerical errors. Data can be exported into most data analysis applications retaining the hierarchical structure. Unlike with Excel, there is no limit to the number of records that can be stored in MS Access. However, the data table needs to be less than two gigabytes, which is largely sufficient for CFSVA surveys. The strength of the Access data entry programme is that it is user-friendly, as it shows the image of the questionnaires (graphic water mark) so that data entry operators can enter the information in the same manner in which it appears in the paper questionnaires. A data entry network can be built that centralizes all records entered by operators into a single computer that the data entry supervisor can use to randomly check questionnaires.

SPSS

SPSS Data Entry Builder allows the design of surveys and forms using the drag-and-drop interface and a library of sample questions that one can edit. One can also create new questions not in the library. In SPSS forms, you can embed rules and filters, and validation procedures. The SPSS data entry platform can also automatically create data files and dictionaries that can be immediately imported into SPSS for analysis.

Epi Info 3.x

Epi Info is a Windows-based programme developed by the U.S. Centers for Disease Control and Prevention. It is freely available on the web at www.cdc.gov/EpiInfo/, which also offers online support. It is one of the most widely-used applications for anthropometric data analysis. Epi Info allows the data entry operator to customize the data entry mask and control for categorical variables and filters. The earlier version of Epi Info (EpiInfo 6.0) was a DOS-based programme. In 2000, the CDC developed a Windows-based Epi Info. Although there were bugs in the earlier Windows-based version, the more recent releases function well.

Epi Info works closely with MS Access as it creates the latter's (.mdb) databases. However, to run Epi Info, it is not necessary to install MS Access, nor does the operator need to know how to work in MS Access.

CSPro

Census and Survey Processing System (CSPro) is a Windows-based data entry and editing programme developed by the U.S. Census Bureau. CSPro has the ability to support the development of effective data entry and the editing of programmes for complex national household surveys. The application is easy to acquire and to learn. Anybody with a basic understanding of databases can learn how to use CSPro in two weeks. It is a free data entry software (available at www.census.gov), and has the added benefit of exporting data easily to a wide variety of analysis software formats.

NutriSurvey

The main purpose of NutriSurvey is to integrate all steps of a nutrition baseline survey into a single programme. The strength of the software is in preparing a suitable questionnaire for entering data and evaluating the results. The programme's standard nutrition baseline questionnaire can easily be customized for a specific survey. For further statistical analysis, the data can be exported to SPSS or another statistical programme.

Microsoft Excel

The Excel application is usually included in a Microsoft Office software package. This easy-to-use spreadsheet programme works with two-dimensional data. However, with Excel it is not possible to develop one-to-many relationship links within household units in one spreadsheet. Likewise, the maximum number of records allowed in Excel is 65,536, and the programme can accommodate only 255 variables in one spreadsheet. Moreover, data masks cannot be developed to minimize entry-level errors. Hence, for complex surveys, such as the CFSVA, the use of Excel by enumerators will increase the probability of data entry errors and inconsistencies.

Table 4.21: Summary of data entry platforms

Software	SPSS	Epi Info	CSPro	Nutri-Survey	MS Excel	MS Access
Type	Commercial (if special option for data entry)	Freeware	Freeware	Freeware	Commercial (usually part of the Windows Office suite)	Commercial (usually part of the Windows Office suite)
Availability	Fair	Excellent	Excellent	Excellent	Excellent	Good
Ease of use (creation of data entry programme)	Fair	Fair	Fair	Fair	Good	Poor
Ease of use (data entry)	Good	Good	Good	Good	Fair	Excellent
Data output	All formats	All formats	All formats	All formats	All formats	All formats
Multi-user data entry	Yes (requires second licence)	No	No	No	No	Yes
Customized data entry forms	Yes	Yes	Yes	No	No	Yes (with graphic watermark)

4.3.3.3 Overview of MS Access data entry mask

The MS Access data entry tool contains a "front-end/back-end" configuration. Briefly, the front-end is the data entry template containing the forms and control code for data entry; it displays the image of the actual questionnaire (see: Image Watermark in Annex 9), with each question associated with a box where the data entry operators enter the answer. The front-end configuration is installed on each of the computers to be used for data entry.

Conversely, the back-end is where the data tables are saved and are linked to the front-end. The back-end can exist either in each data entry operator's computer or in a single computer when using a network; in the latter case, only the supervisor has access to the back-end. Microsoft provides a detailed explanation of the steps involved in creating a front-end and back-end. Visit the following URL for more assistance: **http://support.microsoft.com/?kbid=304932.**

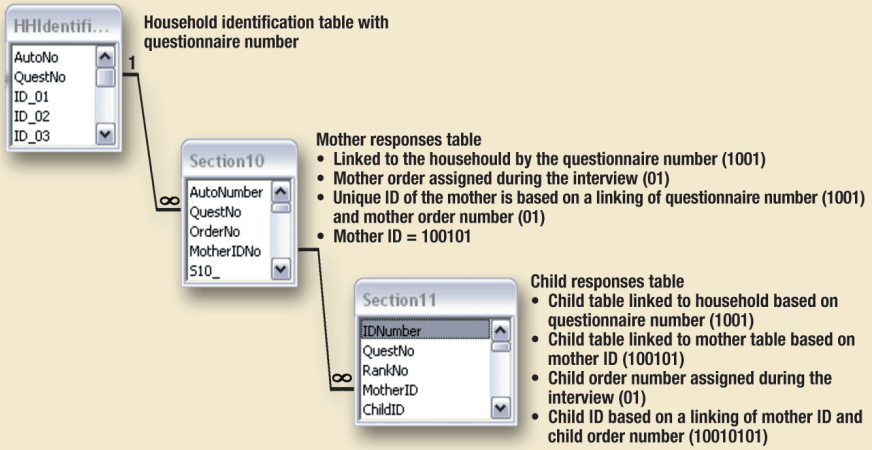
A computer network is a setting in which the front-end is installed in each computer to be used by the data entry operators, and the back-end is installed only in the supervisor's computer. This back-end is meant to centralize all entries from the many data entry operators' computers. Setting up a computer network is a convenient way for the data to be automatically merged into the different tables in a single computer. Also, it allows the supervisor to have control over the entered data and limits data entry operators' ability to manipulate data once it is stored.

In building data entry masks using MS Access, it is critical to know the following key characteristics of the application in order to prevent entry errors:

1. Multiple Related Tables

Microsoft Access allows a maximum of 255 variables in one table. However, a standard CFSVA may have more than 500 variables. Therefore, household data will be stored in several tables linked together (in a one-to-one relationship) by their unique household ID. Since CFSVAs collect demographic data for all household members and anthropometric data for the members in the reference age group, it is necessary to record data by member as well as by household. As demonstrated in Figure 4.6, the relationship between the household, mother, and child is based on the questionnaire number and the order number of the mother and child.

Figure 4.6: Interlinked tables in MS Access



When creating the data entry application with sub-forms, keep in mind that it is easier for the data entry operator to enter the data if the data entry page contains only those questions that are on the page of the actual questionnaire. In other words, the data entry page should mimic the questionnaire page exactly.

2. Location Codes

The analyst may want to classify households based on their location, in which case a location code is essential. Although enumerators will record the cluster or village ID on the questionnaire, data entry operators must be allowed to enter only the valid location codes selected for sampling. This will minimize entry error. Annex 9 offers a detailed discussion on creating data entry masks to minimize entry errors. Moreover, as the names of the administrative levels of the cluster are already defined, it is not necessary for the data entry operators to record the type of location (e.g. region, province, district, or village), as this will only slow down entry speed.

3. Limiting Data Entry

A key role of a data entry tool is to minimize entry errors by controlling for impossible and unlikely responses. The data entry mask in Microsoft Access can control impossible values and outliers, flag possible data entry errors during the entry process, and block the entry of values if a filter question for those values has previously been asked. Depending on how significant the error is, the application can alert the entry operator, restrict the values that can be entered, or refuse the entry of a value. “Smart” programming is particularly important in entering a large amount of survey data. The five main types of controls that can be employed when developing a data entry mask are discussed in Annex 9. Although the examples are not exhaustive, the types of solutions they offer can be applied to other situations.

4. Handling Missing Data

For numeric data, missing values can mean that either the answer is “0” or “not applicable” or the data was not available at the time of the interview. For example, if a household cultivated maize in the past year but could not report the total expenses involved in maize cultivation, then the data is unavailable. If the household did not cultivate maize in the past year, then the question of “total expenses for the activity” does not apply to the household. In this case the response recorded on the questionnaire should be “not applicable.” If, on the other hand, the question is about how many shovels or vehicles the household owns, then a blank response may mean that the household does not own any of these items.

The person handling “missing data” should do so on a case-by-case basis, carefully reading through the completed questionnaire to understand why the data is missing and how to handle this. A well-designed questionnaire will typically assign codes for “not applicable” cases, and for household answers of “I do not know.” (Refer to the example for filters in Annex 9.)

4.3.4 Data entry

Once the data entry programme is written, it can be used to enter data. Data entry is a very important step in the data management process, and should be carefully

monitored. As with data collection, usually many people are involved in data entry, and the quality of the final data set depends on the quality of their work.

4.3.4.1 Testing the data entry programme

Testing the data entry programme is an extremely important step, to be undertaken prior to actual data entry. This should be done with filled-in questionnaires. The entered data should then be cross-referenced with the questionnaires to make sure the data in the fields are appropriately ordered. Once the data entry is declared to be error-free, the data entry process can begin.

4.3.4.2 Managing data entry operators

Managing the data entry operators is a key step to ensuring an accurate database. Continuous monitoring of data entry work from the beginning to the end can significantly reduce entry errors. The time it takes to completely enter the data from a household survey depends on a number of factors, including the design of the data entry application, the number of variables in the questionnaire, design of the questionnaire, the key stroke speed of the entry operator, the number of operators involved in data entry, and the overall management of the entry process. Key steps in managing the process are.

a. Hire entry operators with prior experience of data entry. Familiarity with Microsoft Access is an added advantage but is not required. Key stroke speed, attention to detail, common sense, initiative for solving problems, and a willingness to seek guidance when necessary are all important qualities of a good data entry operator.

b. Provide training to data entry operators on the questionnaire, survey objectives, and data entry application (much like enumerator training). The training should allow the operators to enter data and identify mistakes. Keep in mind that data entry operators will be better able to identify and solve errors if, in addition to the data entry template and data masks, they understand the survey, its objectives, and the meaning of the questions. One day of training is usually sufficient, followed by a half-day to one full day of practice data entry, where the supervisor reviews the quality and accuracy of their work.

c. Remember that it is normal for data entry operators to work slowly at first. It is critical to focus on accuracy rather than speed. As the operators get to know the programme, the questionnaire, and the typical responses, their speed will increase. The daily target for amount of data entered should be set only when the supervisor is pleased with the quality of work after several days of data entry.

d. It is good practice to assign a particular computer to one (or two, if they are working in shifts) data entry operators. This way it is easier to monitor the quantity and quality of each operator's work.

e. To ensure the accuracy of entered data, the supervisor should be responsible for checking a certain number of questionnaires per day (usually about 5 to 10 percent of all questionnaires) against the entered data.

f. Keep a daily record of how many questionnaires per day each data entry operator has completed. It is advisable to pay the entry operator by day rather than by

page or by questionnaire, as the latter discourages the operator from taking time to solve any problems he/she encounters, and could increase the number of errors in the final database. However, after the first or second day of data entry, it is a good idea to set daily quotas for questionnaires entered in the database by each data entry operator.

4.3.5 Paper questionnaire management during data entry

Tracking the filled-in questionnaires is extremely important during the entire survey process, ensuring that none is lost or entered more than once.

a. The filled-in questionnaires should be organized by cluster, in boxes or in envelopes. Clearly write the name of the cluster on the envelope or box. Identify an area where questionnaires will be stored prior to data entry, and a second area, clearly separate from the first, where questionnaires will be stored post–data entry. Identify a third area, near the data entry computers used, where the envelopes/boxes of questionnaires currently being entered can be stored. Finally, identify a fourth area where questionnaires with problems can be stored until their problems can be resolved.
b. Prepare a register in which all of the clusters and the numbers for the filled-in questionnaires in each of the clusters are recorded. This can be done by recording the name of the cluster and the questionnaire numbers contained in that cluster (e.g. 130052 through 130075). When the data entry operator selects a cluster for entry, she/he should sign and date the register to assume responsibility for that cluster's questionnaires. This will also help to track the questionnaires (which are also all clearly numbered).
c. Each data entry operator should work on one envelope of questionnaires at a time, and should be responsible for all of the questionnaires in that envelope. As data entry for each questionnaire is completed, the questionnaire should be returned to its envelope/box. This will prevent the loss or misfiling of questionnaires.
d. When the data entry operator starts to enter information from a questionnaire, he/she should record his/her name and/or code on the cover page of the paper questionnaire. If desired, this code can also be entered into the database during data entry, allowing easy identification of data entry operator during analysis.
e. When all information from a questionnaire is entered, a clear mark (e.g. a large checkmark or a slash) should be made across the entire front page using a highlighter. Once all questionnaires in an envelope/box are entered, the same mark should be made on the envelope/box.
f. When the data entry operator puts the envelope/box containing the already entered questionnaires in its designated area (see point a.), she/he should sign and record the date in the register.

4.4 HOUSEHOLD DATA ANALYSIS AND PROCESSING

4.4.1 Objective

This section is designed to help analysts (who should already have an in-depth knowledge of statistics) to analyse the household survey data generated by CFSVAs. These guidelines do not elaborate on common statistical or data management techniques, as this information is beyond the scope of these guidelines and must be

acquired through academic course work and/or on-the-job training and supervised experience.

Why analyse primary data?
The CFSVA Food and Nutrition Security Conceptual Framework describes how various factors influence the food security situation and vulnerability of households. Using information obtained from various sources, the analyst describes and evaluates household food security status, the factors that influence household food security, the livelihood strategies employed, and the health and nutritional status and other livelihood outcomes at the household level.

Information generated by the CFSVA is used to explain how different households are exposed to risk and how they manage to cope. This information is combined with data obtained from secondary sources to describe the geographic, economic, and social context and explain the risk factors that influence the extent of vulnerability and the capacity to cope with shocks.

4.4.1.1 A note on statistical software
Because WFP uses SPSS for most of its quantitative data analysis, the guidance presented here focuses on that programme. However, experienced statisticians may choose to use other software packages.

For most cluster analysis, and often for principle component analysis, WFP-VAM uses ADDATI,[62] but SPSS can process this analysis, too.

For anthropometric z-score calculations of under-5s (stunting, wasting, underweight), WHO ANTHRO 2005[63] is used. Epi Info is essentially obsolete unless the ENA add-on for EPI Info is used.

4.4.2 Preparation for the analysis

4.4.2.1 Hierarchical data structure
CFSVAs consist primarily of household data. However, information gathered at the household level often includes data on individual household members, such as age, sex, children's education, nutritional status of mothers and children (under 5), women's childcare practices, and women's knowledge of HIV/AIDS. This may result in multiple "units," or cases, from each household. Additionally, data may be gathered at the village or community level that is pertinent to each household in the community (such as presence of schools and health clinics).

These data need to be organized into several data files, one for each unit of analysis, corresponding to the level at which the data were collected. For example, the member-level information (e.g. age, sex, children's education) should be saved into one file, while household-level information (e.g. assets, expenditure, current food consumption) should be saved into a separate file. There should be a separate file for anthropometric

62. This software can be downloaded for free at: http://cidoc.iuav.it/~silvio/addawin_en.html
63. This software can be downloaded for free at: http://www.who.int/childgrowth/software/en/

information for children (sex, age in months, height, and weight). Meanwhile, a different file should be used for child-care data. Similarly, if village-/cluster-level information is collected, a separate data file would be needed.
For CFSVAs, data can generally be organized in up to five data sets:
1. Village
2. Household
3. Individual
4. Mother
5. Child

It is essential to develop a data management plan before data entry.[64] Data entry application may automatically produce the five data sets (if designed to do so) or one large data set that needs to be reorganized into several data sets.

To obtain information about each of these levels, each of these data sets needs to be analysed. However, the analyst may desire to combine information from different data sets into one combined data set. Using SPSS, queries cannot be made between individual data sets; therefore, data sets must be merged in SPSS using a "many-to-one" relationship.

The merging of different data sets needs to be done so that member-level data sets can add to the information gathered from the household. Data analysis should in general be done only at the lowest level contained in that data set (i.e. member level), as described in Box 4.8.

Box 4.8: Relating child nutrition with other indicators

An analyst wants to look at child malnutrition as it relates to the household water source. There can be multiple children in each household, so the relationship between these two data sets (child and household) when merging is many children to one household. Using SPSS, this means merging the household data set into the child data set. This preserves all child information and keeps the number of children in the data set the same. However, in this merging process, some household information is lost (e.g., those with no children under 5) and some is duplicated (e.g. where there is more than one child in a household).

The resulting data set is used only for child-level queries (e.g. to answer the question "What percentage of wasted children live in households with unsafe drinking water supply?")

This merged data set cannot answer the question "What percentage of households have an unsafe water supply?" because some households were duplicated (i.e. those with more than one child) and others were deleted (i.e. those without any children). This question should be analysed within the household data set.

This merged data set cannot answer the question "What percentage of households have a wasted child as one of the members?" Analysis in this direction (from higher to lower aggregation level) is uncommon, and generally not recommended. To answer such questions, merging in an alternate direction would be required.

64. See Section 4.2.5

4.4.2.2 Organization of the database

Organizing the database is an important step to getting a clear idea of the variables the analyst is going to consider. It is also helpful to manage and analyse the data by different individuals. **It is a good practice to make a copy of the database and keep it in a separate folder.** The following are key aspects of database organization:

a. Verify that all **variable names** clearly identify the question in the questionnaire. This can be easily done by using the question's code. Do not change the variable names unless it is absolutely necessary. Changing a name may complicate the identification of the particular variable when comparing it to the original raw data (from MS Access or another data entry tool), especially for other analysts who might access the data later. Additionally, if additional cases are to be appended to the database, differing variable names will impede the process.
b. Often variable names are cryptic; therefore it is necessary to enter **variable labels** to clarify what each variable is. A well-designed data entry programme, properly exported to SPSS, will already have appropriate labels for all variables, but this should be carefully checked. If the labels are clearly written and correctly spelled, it will be easier and quicker to create tables for reporting.

Box 4.9: Example of variable names and labels

In the questionnaire:

HQ5.1b What is the main source of drinking water for your family? 1…, 2…, 3…

The variable name could be HQ5.1b.
The variable label could be "drinking water source."

c. The variable type should also be correctly identified (usually "string variable" for letter/word values, and "numeric variable" for numbers). For categorical variables, it is necessary to enter the value labels for each variable, following the coding from the questionnaire. This information is essential for analysing the data and also for cleaning the categorical variables in the data set (see section 4.4.2.3 on data cleaning).
d. Identifying the measure (scale, ordinal, or nominal) of a variable is a key part of database organization. This information enables the software to conduct appropriate analyses with specific variables.
e. During data cleaning it is also important to specify whether a variable has one or more missing values. It is not uncommon for an analyst to be the first person to discover that a variable is missing values. Coding missing data can be done in several ways. However, each variable, and possibly even different analyses of the same variable, will have their own specific needs, and so there is no cut-and-dried rule for dealing with missing data.
f. Another step involves data recoding. This process is particularly useful for categorical variables. For example, yes/no questions are best coded as 1/0 (not 1/2). The 1/0 option is preferred because a simple mean illustrates the frequency, and in case of regression analysis, the yes/no questions are already recoded as

binary variables, with the sign of the regression coefficient pointing in the intuitively correct direction. If the data entry programme is well designed, this should not be necessary. Boolean variables from MS Access translate automatically in SPSS into 1 (yes/present/true) and 0 (no/absent/false). A good and essential rule is not to lose data during recoding. For example, do not recode and simultaneously replace a continuous variable with a categorical variable. Instead, keep the original variable and create a new categorical variable.

g. A good practice is to keep only those variables meaningful to the analysis in the final dataset. If the analyst creates too many working variables before arriving at the final working data set, these variables will need to be deleted after computation, otherwise the size of the database will increase exponentially and become difficult to manage.

Figure 4.7 gives an example of all the fields that must be organized before data analysis.

Figure 4.7: Example of a database

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	hhid	Numeric	7	0	household ID	None	None	8	Right	Nominal
2	Date	Date	20	0	Date	None	None	17	Right	Scale
3	IntID	Numeric	4	0	IntID	None	None	7	Right	Nominal
4	IntName	String	200	0	IntName	None	None	17	Left	Nominal
5	SupID	Numeric	4	0	SupID	None	None	8	Right	Nominal
6	State	Numeric	2	0	State	None	None	6	Right	Nominal
7	STATEID	Numeric	8	0		None	None	8	Right	Scale
8	LEVELS	Numeric	8	0	levels of analysis	{1, rural population}...	None	8	Right	Nominal

4.4.2.3 Data cleaning

Data cleaning is an essential step in data analysis. Every dataset contains some errors, and a significant amount of time in data analysis is spent "cleaning" the data.

Data cleaning can commence once the data are organized into different files. Data cleaning does not mean simply confirming that the data recorded on the paper questionnaires is the same as that in the dataset. It also entails several iterative steps of checking the dataset(s) to ensure that the data are credible.

Usually the cleaning of CFSVA data sets is done in several stages. The initial part of the data cleaning can be done with the software used for data entry[65] (most often MS Access). Cleaning should not be done as an automatic process but, rather, as a critical, well-thought-out series of recorded decisions.

UNIQUE ID

The first step in the data cleaning process is to ensure that the total number of households in the data set equals the total number of filled-in paper questionnaires. It is important to carefully review the data sets to confirm that all questionnaires have been entered only once and that all unique IDs are truly unique. This step ensures that data sets can later be merged and other household specific variables added, if necessary.

65. See section 4.3.3.

In the more recent versions of SPSS, there is now an option, in the "Data" menu, to "Identify Duplicate Cases," which makes this process very simple. If this option is not available, running a "frequency of the household ID" can be useful for detecting the presence of duplicate households. Household IDs resulting in a frequency of 2 or more have duplicates.

Additionally, it is a good idea at this point to take a random selection of questionnaires and compare them to the database (this should also happen as part of the quality control step during data entry). This verifies that all questions are being entered correctly and that variables are not being mislabelled.

Check the variables

The next step in data cleaning become more subjective, thus it is important not to make any permanent changes to the data unless you are absolutely confident in the decision. Regularly save backups of the database (without replacing earlier backups) so that any changes made can be undone at any time. These steps include checking for outliers and checking for errors/inconsistencies.

Check for outliers

An outlier is an observation that is numerically distant from the rest of the data. Statistics derived from data sets that include outliers will often be misleading. In most samplings of data, some data points will be further away from their expected values than what is deemed reasonable. In the presence of outliers, any statistical test based on sample means and variances can be distorted. Estimated regression coefficients that minimize the sum of squares for error (SSE) are also sensitive to outliers. Outliers can be caused by data collection/data entry errors or by extreme observations that for some legitimate reason do not fit within the typical range of other data values (High 2000).

Check the distribution of data values by levels of a categorical variable, if available. This procedure should always be one of the first steps in data analysis, as it will quickly reveal the most obvious outliers. For continuous or interval data, visual aids such as a dot plot or scatter plot are good methods for examining the severity of any outlying observations. A box plot is another helpful tool, since it makes no distributional assumptions, nor does it require any prior estimate of a mean or standard deviation. Values that are extreme in relation to the rest of the data are easily identified.

Running a frequency table or simple descriptive statistics could also be useful for detecting outliers. Working with outliers in numerical data effectively can be a rather difficult experience. Neither ignoring nor deleting them is a good solution. If nothing is done with the outliers, the results will describe essentially none of the data – neither the bulk of the data nor the outliers. Even though the numbers may be perfectly legitimate, if they lie outside the range of most of the data, they can cause potential computational and inference problems (High 2000). Outliers are not "missing," just too high or low given our expectations; hence they should not be recoded as missing data. There are a couple of ways to deal with outliers.

- Since means are sensitive to extreme values, median values can be used instead of means.
- Maintain the "raw" version of the data, which retains the outliers, but create a "processed version" in which new variables are created that, for example, replace outliers with medians. Create a variable that denotes whether an outlier has been replaced by a median. That way, no data are discarded.

In both cases, it is crucial to report how outliers were managed.

Box 4.10: Possible effects of outliers

- Bias or distortion of estimates (especially of the arithmetic mean)
- Inflated sums of squares (which make it unlikely to partition sources of variation in the data into meaningful components)
- Distortion of p-values (statistical significance, or lack thereof, can be due to the presence of a few, or even one, unusual data value)
- Faulty conclusions (it is quite possible to draw false conclusions if irregularities in the data have not been investigated)

Check for errors/inconsistencies

Impossible values are often found in data sets, in spite of the filters used in the data entry programmes. Sometimes the values are absolutely impossible or contradictory to the information given in prior questions.

Box 4.11: Example of inconsistent values

- The number of the household members is 50.
- People with little or no land had a considerable harvest.

Once an inconsistent value has been identified, the data should be checked on the paper questionnaire to exclude the possibility of data entry error. If the data was entered wrongly in the paper questionnaire, the analyst should be able to decide whether to exclude the value, based on his/her experience and the type of variable. The rule is to change "absolutely impossible" values into "missing values," that is, if there is no way to determine if it is too time-consuming to investigate for the correct value. However, this is a subjective choice and should be approached with absolute caution.

Sometimes, even if the answer appears clean when you compare it with another variable, there is still an evident contradiction. In this case, the rules for an efficient data cleaning will be first to check the original questionnaire; if the answer is not there, look at other variables that can have a connection to those contradictory variables. If even this solution does not yield results, record the value of both the variables as missing.

Box 4.12: Example from Laos database

In the child database there is information about child demographics and household demographics.

Section 1, question 1.7: the enumerator should complete the information about the household demographics (number of people in the different sex and age groups).
Section 10, questions 10.3 and 10.4: the enumerator entered the information about the child's age and sex. In the database, the analyst found many inconsistencies that were difficult to solve, including households where the number of people in a specific age group does not match the number of children measured.
In this situation, the analyst, after cross-checking with the paper questionnaire, should try to find the truth in other variables (e.g. by looking at the variables related to education to see if the child was included in the wrong age group) or exclude the case from the analysis.
In other cases, the sex of children is different in the two sections. In this case, the information in Section 10 should be more accurate because the children were present during the measurements. So the analyst changed the variable 1.7 based on the information collected in 10.4.

Usually a questionnaire is developed with the flow of the questions kept in mind. In many cases there are skips in the questionnaire that allow the interviewer to bypass questions not applicable to a particular respondent. The data should be entered accordingly. For example, if a household did not cultivate any land, questions regarding harvest and crop types are not applicable to them. However, a well-designed data entry programme should automatically skip the fields that are not applicable to a household based on the information entered for the filter question.

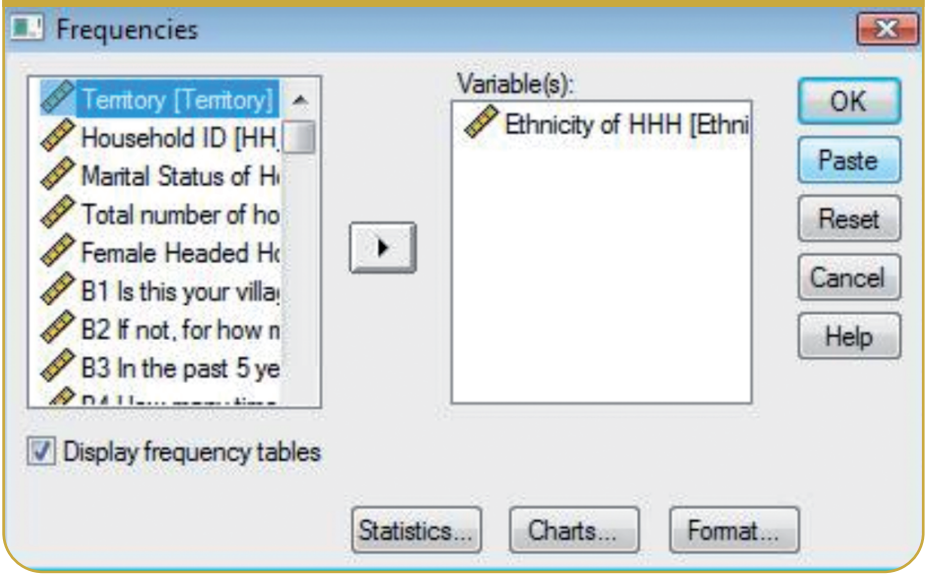
4.4.2.4 General rules for data cleaning

- Do not start guessing, predicting, or assigning values. Even if a value seems obvious, do not make a change unless it is supported by clear evidence and the change is recorded.
- Bear in mind that most often the database will be managed/analysed by different people after the cleaning. Prepare an easy-to-read, clean database in which all the variables have their basic information; this will reduce mistakes and minimize time spent trying to comprehend the data set.
- Save a copy of the unchanged database before making any changes to it.
- Recode the history of changes in a syntax file. A record of the changes will be invaluable and can help replicate the same cleaning in different backups.
- In case of contradiction/inconsistency/doubts in the database between variables:
  1. Check the original questionnaire.
  2. Check the validity of the data by comparing them to other variables in the database.
  3. Change the value only if you are 100 percent sure.
  4. Consider the case as missing data.
  5. Keep the syntax for recoding automatic changes (if any).

4.4.3 Data analysis

4.4.3.1 Standard practices

Figure 4.8: Pasting syntax



This section discusses several standard practices that are useful in CFSVA data analysis. By following these general guidelines, CFSVA analysts will have more compatible methods of organizing their analyses.

Use of syntax

In SPSS, the syntax is a log of all transformations and procedures used. Syntax can easily be generated in SPSS. Using syntax is a matter of personal experience and preference, ranging from only minimal use for data cleaning to use for conducting all analyses.

In most of the interactive menu for a transformation or procedure in SPSS, there is a "Paste" option. This will save the syntax of the command or transformation in the **most recently** opened syntax file (see Figure 4.8).

It is recommended to keep several syntax files: one for data cleaning, one for key complex transformations, and others for the main analytical steps. Keeping the syntax of transformations and key analysis is good practice. That way, if you are ever uncertain about how a variable was created or corrected, you can go back and check.

Syntax can be copied, pasted, and edited in the SPSS Syntax Editor. It is a good practice to write a brief description of the procedure before each syntax command. Separating different parts of the analysis is also considered good practice.

Advantages of using syntax

There are several reasons for using syntax. It can improve efficiency and transparency, and also save time.

A lot of transformations for different variables are similar. A clever use of the Copy/Paste and eventually the Find/Replace function in the Syntax Editor can avoid repetitive chores, save time, and even reduce the chance for error in transformations. Some analytical procedures might take more manipulation than others. This can easily be done by using the interactive menus in SPSS. However, if an analyst wants to change the procedure after the analysis, he/she will have to rebuild the entire procedure, reset all options, and reselect all the variables. Clearly it takes a lot of unnecessary extra time before the same procedure with the same results can be reproduced. If the syntax is saved, however, no time is wasted at all.

At a later date, the analyst or a colleague might want to review the transformations performed or the exact way a procedure was conducted. Use of syntax will enable a review of the draft analysis and make necessary corrections before finalizing the report. It is not uncommon to identify inconsistencies in the data or problems in previous procedures. Moreover, others should be able to see exactly how new variables were defined. Using syntax will give instant access to the formula used to create the new variables. It is recommended to keep a complete syntax of key transformations and procedures.

Data backup

Frequently save backups of the data set. An erred transformation of a variable or a manipulation of the data file can result in lost data. Each successive version of a database should be independently saved. This allows the retrieval of original data if such a mistake occurs.

Labelling variables and values

The cleaned database should have complete variables and values recorded. When creating new variables, be sure to enter suitable variable labels and values of categorical variables. This will enable another analyst to easily and quickly interpret the variables in the data set. As described in section 4.4.2.3, it is recommended that the variable names reflect the questionnaire number, and the variable label provide a more detailed and widely understood name. Calculated variables should have an appropriate name, and a label that clearly identifies the variable to future analysts.

4.4.3.2 Types of variables

The different types of data collected in a household survey, including age, sex, income, assets, and names of districts/provinces, can be categorized according to their measurement scale. Four measurement scales are generally used in statistics: nominal, ordinal, interval, and ratio. Nominal and ordinal variables are considered to be **categorical variables**, while interval data and ratio variables are considered **continuous variables.**

Categorical variables

A categorical variable is one for which each response can be put into a specific category. The categories are usually labelled and coded. These categories must be

both mutually exclusive and exhaustive. Mutually exclusive means that each possible survey response should belong to only one category, whereas, exhaustive requires that the categories cover the entire set of possibilities. If the age categories are 0–6, 7–12, 13–18, they are mutually exclusive, as a person can never be in two of these categories at the same time. To be exhaustive, we have to add a category ("19 or more") so that all possible cases are covered. Categorical variables can be either nominal or ordinal.

Nominal: A nominal variable describes a name or category. Contrary to ordinal variables, there is no "natural ordering" of the set of possible names or categories. Sex, household status, and type of dwelling are examples of nominal variables. Another example is type of crop, which could be categorized as 1 = wheat, 2 = rice, 3 = maize, 4 = sorghum, 5 = millet, 6 = other. Nominal variables cannot be analysed using means; the mode can be used.

Ordinal: Ordinal variables order (or rank) data in terms of degree. Ordinal variables do not establish a numeric difference between data points. They indicate only that one data point is ranked higher or lower than another (Shawna, J. *et al.* 2005). They are not customarily analysed using means. The variable "food consumption group," for example, is ordinal because the category "acceptable food consumption" could be considered better than the category "poor food consumption." There is some natural ordering, but it is limited since we do not know by how much "acceptable food consumption" is better than "poor food consumption."

Example: Variable: Food consumption group, where:

1 = poor food consumption
2 = borderline food consumption
3 = acceptable food consumption

Binary: A special type of categorical variable is a dichotomous/binary variable, which is a nominal variable consisting of only two categories (or levels). Observations can be classed into two groups: male/female, group 1/group 2, true/false, yes/no. All cases having a certain characteristic. For example, "household has female head" could be coded with a value 1, for "yes", while all other cases without that characteristic could be coded with a value 0, for "no". Coding 1/0 allows calculations, which are normally not possible with a nominal variable. For example, the percentage of cases having the given characteristic (e.g. percentage of female household heads) corresponds with the average of the variable.

Continuous variables

A variable is said to be continuous if it can take an infinite number of real values. Continuous variables are interval or ratio variables.

Interval variables

The numbers assigned to objects have all the features of ordinal measurements, and in addition, equal differences between measurements represent equivalent intervals. That is, differences between arbitrary pairs of measurements can be meaningfully

compared. Operations such as addition and subtraction are therefore meaningful. The zero point on the scale is arbitrary; negative values can be used. However, ratios between numbers on the scale are not meaningful, so operations such as multiplication and division cannot be carried out directly. For instance, the phrase "today it is 1.2 times hotter in degrees Celsius than it was yesterday" is not very useful or meaningful; in degrees Fahrenheit it might be 1.4 times hotter. Stating that the birth year of person A is 5 percent higher than the birth year of person B is also not useful or meaningful.

The central tendency of a variable measured at the interval level can be represented by its mean, median, or mode, with the mean giving the most information. Variables measured at the interval level are called interval variables, or sometimes scaled variables, though the latter usage is not obvious and is not recommended. Examples of interval measures include temperature in Celsius scale or Fahrenheit scale.

Ratio variables

A ratio variable, has all the properties of an interval variable, but also a clear definition of 0.0. When the variable equals 0.0, there exists none of that variable. Variables such as height and weight are ratio variables. Operations such as multiplication and division are therefore meaningful. The zero value on a ratio scale is non-arbitrary. The central tendency of a variable measured at the ratio level can be represented by its mode, its median, its arithmetic mean, or its geometric mean, as with an interval scale.

Table 4.22: Summary of statistical measures by type of variable

Okay to compute	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Median and percentiles	No	Yes	Yes	Yes
Add or subtract	No	No	Yes	Yes
Mean, standard deviation, standard error of the mean	No	No	Yes	Yes
Ratio, or coefficient of the variation	No	No	No	Yes

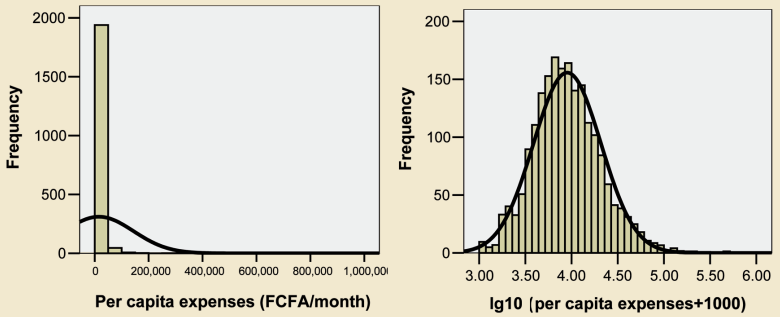
A continuous variable can be categorized into a categorical variable to facilitate months: 24–35 months, 36–47 months, and 48–59 months. The actual age of the children collected in a survey can be categorized into these groups by giving a code such as 0–5 months = 1, 6–11 months = 2, and so on.

Distributions of continuous variables

The distribution of continuous variables should be considered during analysis. Most common procedures and statistics assume that variables are normally distributed. When the distribution of a continuous variable is highly non-normal, alternative data summaries should be used (e.g. median and not mean), and tests of significance that do not assume normality should be used (non-parametric tests). Sometimes, variables

can also be transformed to have a more normal distribution. This transformed variable can then be used, for instance in a regression or a principal components analysis (PCA).

Figure 4.9: Distribution of per capita expenditure



Expenditures, revenue, and production typically have a skewed distribution, as in this example from the Cameroon CFSVA. The mean is 15,500 FCFA/month per capita, however, a few households have monthly per capita expenditures reaching 500,000 to 1,000,000 FCFA/month.

Using the per capita expenditures in a regression model or correlation, or running a t-test or ANOVA to compare means, would violate the assumption of normal distribution. Hence a logarithmic transformation was applied, and the resulting distribution resembles the normal. The transformed variable can now be used in correlations, regressions, PCA, etc.

Constructing indicators: Transformations

Simple or complex transformations are needed to create many of the key CFSVA indicators. Mathematical operations in the compute procedure, categorizing of values in the recode procedure, and other transformations possible in SPSS allow for the combining of different variables or the reconfiguration of variables into desired indicators.

Changing values of a variable

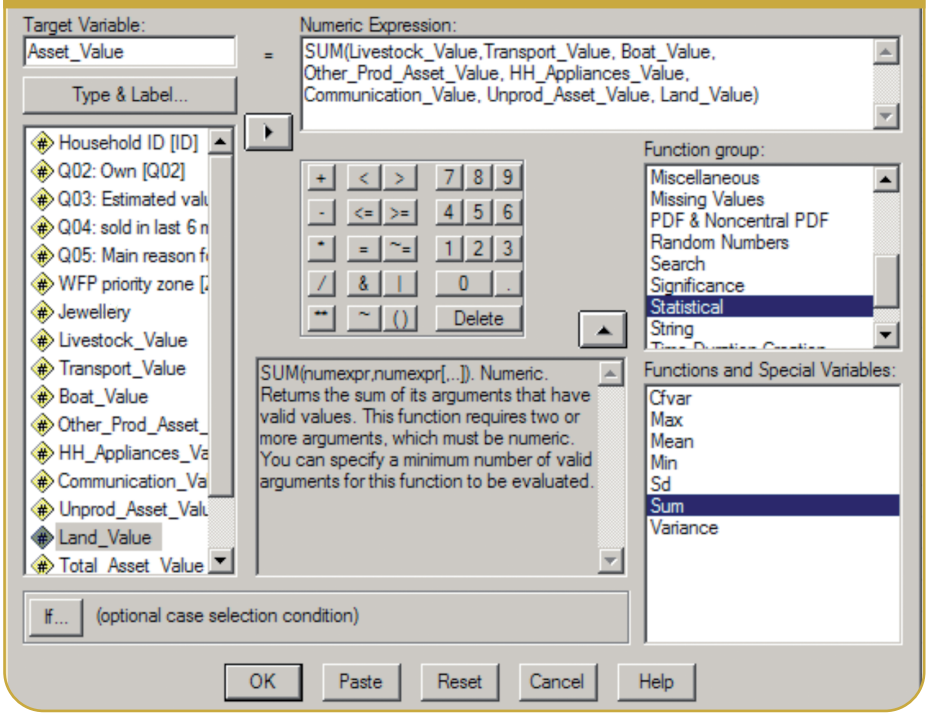
The "Recode" command in SPSS is generally used to change the values in a variable. Variables can be recoded into the same variables or into different variables. Generally, it is recommended that recoding be done into a different variable so that the original data is not lost, particularly in the case of an error in recoding. This command can be used to recode one categorical variable into a new categorical variable, or one continuous variable into a new categorical variable.

It is worth aggregating categories or transforming a continuous variable into a categorical when the list of possible answers is very long (e.g. the relationship of the HH members with the HH head), and/or when some answers have been chosen by few households.

Calculating new variables

New variables can be calculated using the "Compute" command in SPSS. To help compute new variables, SPSS has a number of mathematical operations, including addition, subtraction, multiplication, and division. However, the "Addition" command does not work if the variables added contain missing values. Using the "Sum" command (Figure 4.10) addresses the problem.

Figure 4.10: Computing a new variable using "sum" command



If the analyst adds variables that contain missing values, it is recommended that he/she not use the "Addition" command (+). If it is used, the sum will have a missing value every time there is a missing value in one of the added variables. The command "Sum" would treat all the missing values as if they were "zero," thus not increasing the number of missing values in the variable "Sum."

"Logical operators" can be used to set up conditions ("If" command in SPSS) to create a new variable. Box 4.13 includes a list of the logical operators available in SPSS.

With careful use of these operators, new variables can be constructed for CFSVA analysis. For example, CFSVAs often collect age of children, which is a continuous variable. However, analysts typically create age categories to generate cross tables with other variables like enrolment or drop out.

Box 4.13: Commonly used logical operators in SPSS

<	less than
>	greater than
<=	less than or equal to
>=	greater than or equal to
=	equal to
~=	not equal to
|	or
and	logical "and"

Box 4.14 demonstrates the use of logical operators to create age groups from a variable called "childage."

Box 4.14: Use of logical operators in SPSS

```
If (childage <= 7) agegroup = 1
If (childage >7 and childage <= 12) agegroup = 2
If (childage >12) agegroup = 3
```

However, be careful of missing value codes. For example, if "no answer" is coded as 99, this could be miscategorized as 3, meaning a child more than 12 years old.

The compute command can also use more advanced mathematical functions. For instance, the square root or the logarithmic transformation could be used to normalize a skewed distribution, and TRUNC or RND can be used to categorize continuous variables.

Box 4.15: Some useful mathematical functions in SPSS

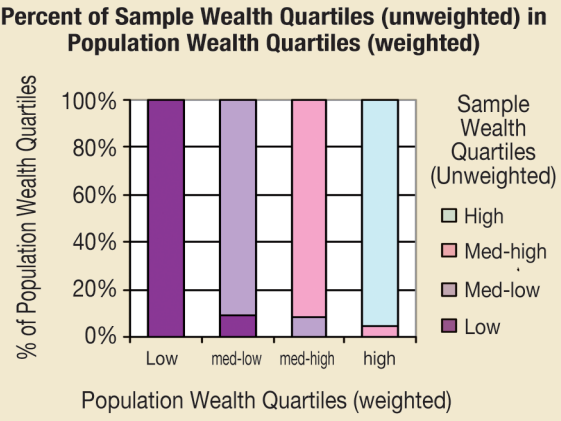
ABS(var)	the absolute value of a variable ABS 13.8=13.8; abs(-13.8)=13.8
RND(var)	the rounded value of a variable: RND(13.8)=14
TRUNC(var).	the truncated value of the variable: TRUNC(13.8)=13
SQRT(var)	the square root of the variable
lg10(var)	the base 10 logarithm of a variable
ln(var)	the natural logarithm of a variable
exp(var)	e raised to the power of the variable

Calculating n-tiles

N-tiles (usually quintiles) can be calculated automatically using SPSS. Under "Transform," use "Rank cases." Under "Rank types," select "n-tiles," and indicate the number of tiles desired, then continue. To deal with tied ranks, usually the mean is used (under "Ties," select "Mean"), although there may be circumstances where other methods should be used.

Most CFSVA data sets use probability weights, hence n-tiles should be calculated with weights on, so that (in the case of quintiles) 20 percent of the weighted sample lies in each quintile, allowing the quintiles to be applied to the population they are representing.

Figure 4.11: Example of incorrect classification of HHs into wealth quartiles



Based on various assets, a wealth index was calculated in the Cameroon CFSVA. When creating quartiles of this index, using the sampling weights, the results can be extrapolated to 25 percent of the population being classified in each quartile. If, however, the weights are turned off when creating the quartiles, 25 percent of the sample household are classified in each "sample wealth quartile," which does not coincide with population quartiles.

The example from the Cameroon CFSVA shows that if the weights are not turned on, 6 percent of rural households are classified in the wrong category.

COMPUTING RATIOS

At household level

Ratios are simply calculated using the "compute" function. Particular care must be taken with 0 values (division by 0 generates a missing value), and that missing values (such as 99,888) are coded and recorded as "missing" in the variable view.

At aggregate level

Although most ratios can be calculated at the household level, it makes more sense for some to be computed at the aggregate level. The enrolment rate at the household level, for example, is often 1 or 0 or missing (when there are no school-age children), whereas the enrolment rate of an entire subgroup is more meaningful. Remember, a statement for a ratio calculated at the household level is different from the same statement for a ratio calculated at the aggregate level, and should be reported as such.

Box 4.16: Example of aggregate ratio vs. household ratio

Calculating dependency ratio in both ways:
Household A has 4 children and 2 productive adults. The dependency ratio is 4 to 2, or 2.
Household B has 1 child, 1 elderly person, and 8 productive adults. The dependency ratio is 2 to 8, or 0.25.
The AVERAGE household dependency ratio = ((2 + 0.25)/2) = 1.125
The AGGREGATE dependency ratio = (sum of all children and elderly/sum of all adults) = (6/10) = 0.600

In order to calculate a ratio at the aggregate level, the values of the variables for the denominator and the variable for the numerator are first aggregated. For example, if calculating school attendance, the sum of all the children attending school in the sample needs to be calculated, and then the sum of all the school-age children needs to be calculated. Then the ratio for that level of aggregation is calculated, using the aggregates in the denominator and numerator. Alternatively, the average for the sample of the numerator and the denominator can be calculated independently, and then these averages are divided to achieve the aggregate ratio. The confidence interval of certain rates can be wide, especially for subgroups. It is therefore important to estimate the error correctly.

Calculation of anthropometric indicators (z-scores)

Epi Info, from the CDC, and Anthro, from WHO, are the two most commonly used applications for anthropometric data analysis. The current standard recommendation for CFSVAs is to calculate and report under-5 anthropometry (stunting, wasting, underweight) using both the NCHS and the new WHO references. Analysis of nutrition should then use the WHO reference data. Only the software WHO Anthro 2005 can currently calculate z-scores using both reference scores.

4.4.3.3 Descriptive statistics

Descriptive statistics are those used to describe characteristics of a sample or population, and they involve exploring the distribution of one variable (frequency) or the distributions between two or more variables (cross tabs). In SPSS, descriptive statistics can be most easily produced by the Frequencies command. Together with simple graphics, they form the basis for virtually every quantitative analysis of data.

Means

a) Simple mean of continuous variables

The (arithmetic) mean is the sum of all the values divided by the number of cases (considering only valid cases and excluding missing cases). It can be used for continuous data (that is, data measured on an interval or ratio scale). The mean is a measure of the variable's central tendency. Statistics such as mean (and standard deviation, defined further on) assume a normal distribution and are appropriate for quantitative variables with symmetric, or normal, distributions. Typical means calculated in CFSVAs include number of household members and monthly income.

b) Value codes of binary variables

While defining the coding of variables, it is recommended to use 1 for "yes" or "present," and 0 for "no" or "not present." In a household survey, the population mean of a variable coded in this way corresponds with the proportion of households that answered "yes" or where the reply was "present," and when multiplied by 100, gives the percentage prevalence of "yes."

Similarly, one could agree to specify 0 for "male" and 1 for "female." However, for the sex of children, to be used in anthropometric data transformations,[66] "male" should always be male = 1 and female = 2. In this case, means cannot be used easily to calculate proportions and percentages.

Medians

The median is the middle of a distribution when the values are ranked from highest to lowest, meaning half of the values are above the median and half are below the median, the fiftieth percentile, i.e. the middle value of a set of observations ranked in order.

66. This is the convention for Epi Info from the CDC and Anthro from WHO.

If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values, unlike the mean, which can be affected by a few extremely high or low values.

The median is a robust statistic, appropriate for quantitative variables that may or may not meet the assumption of normality. It is preferable to use the mean when the data are not normally distributed. For example, some expenditure and income data have very skewed distributions, and the median may be a better summary than the mean. The median can be used with measurement scales that are at least ordinal (that is ordinal, interval, or ratio).

Modes

Modes are rarely used in describing CFSVA indicators. However, they may occasionally be of use. The mode is defined as the most frequent variable value. In the set of values 1, 2, 3, 3, 3, 4, 5, 5, 6, 6, 7, 8, 9, the mode is 3, as it appears more frequently than any other value. There can be more than one mode. Using the mode is more appropriate when there are only a few possible values of the variable (for instance, to describe the household size, the mode could be used). The same is true for categorical variables (for example, "farming" is the most common livelihood strategy of the households.

PERCENTAGES/PROPORTIONS

Frequencies

Frequencies is one of the more common descriptive functions used in the analysis of CFSVAs. They are most commonly used to produce global prevalence (prevalence for the whole data set). SPSS gives two prevalence results: percentage and valid percentage. Percentage includes the missing cases (system missing and those coded as missing) in the denominator, whereas valid percentage excludes the missing cases and includes only the cases with data in the denominator. The analyst needs to interpret which of these to report. If the missing values are assumed to be not different from the values with valid responses, the "valid percentage" statistic should be given; if the missing values are different (nonexistent options, not applicable), the analyst should consider reporting the "percentage" as compared to the total population. Valid percentage is the most common; however, in certain cases percentage may be more relevant.

Box 4.17: Example highlighting the difference in interpretation

In this example from Laos, households were first asked if they cultivated land in the last year. If they responded "no," then the enumerators skipped the rest of the agriculture section, leaving the questions blank. If the response was "yes," then the following question was asked: "What was your main crop cultivated in the past year?" In the data entry, this question on the main crop was left blank (system missing) if there was no response. A frequency of the main crop cultivated results in the following table:

(cont...)

(...cont)

Main crop cultivated

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Glutinous rice	2827	72.0	82.1	82.1
	White rice	365	9.3	10.6	92.7
	Maize	65	1.7	1.9	94.6
	Beans	4	0.1	0.1	94.7
	Cassava	3	0.1	0.1	94.8
	Vegetables	11	0.3	0.3	95.1
	Fruits	19	0.5	0.5	95.6
	Tobacco	1	0.0	0.0	95.7
	Groundnuts and other nuts/seeds	6	0.2	0.2	95.8
	Other	143	3.7	4.2	100.0
	Total	3444	87.7	100.0	
Missing	System	482	12.3		
Total		3926	100.0		

When looking at glutinous rice, two statements can be made. Using the percentage, it can be stated "72 percent of all households cultivated glutinous rice as their main crop in the past year." This statement is of ALL households. Using valid percentage, it can be stated "of the households practicing agriculture in the last year, 82 percent cultivated glutinous rice as their main crop."

Cross tabulations

Cross tabulations are another way of exploring frequencies, and are one of the most common descriptive tools used in CFSVA analysis.

Unlike frequencies, cross tabs include only the valid cases. When calculating prevalence in SPSS, three options are commonly used: percentage rows, percentage columns, and percentage total. This will determine how SPSS calculates the prevalence in each cell. For each cell, the numerator remains constant: the number of valid cases belonging to the two groups. The denominator in percentage rows is the total number of cases in the row cells. The denominator in percentage columns is the total number of cases in the column cells. The denominator in total percentage is the total number of cases in ALL cells, which is equal to the valid number of cases in the data set.

Understanding this difference between percentage rows, columns, and choosing the correct one to report is critical. As a general rule, the "independent variable" should be put in the columns and the "dependent" in the rows; column percentages are more important than row percentages. For instance, we can look at the influence of wealth on food consumption. If we put wealth in the column, we should focus on the column percentages and compare them across the rows (food groups).

In Table 4.23, three food consumption groups were cross tabulated with the quintiles of wealth score (the first being the poorest, the fifth being the richest).

Table 4.23: Example of cross tabulation

			Quintiles of wealth score					
			1	2	3	4	5	Total
Food consumption groups	Poor Food Consumption	% within Food consumption groups	67.1%	13.4%	8.5%	3.7%	7.3%	100.0%
		% within Quintiles of wealth score	7.1%	1.4%	0.9%	0.4%	0.8%	2.1%
		% of Total	1.4%	0.3%	0.2%	0.1%	0.2%	2.1%
	Borderline Food Consumption	% within Food consumption groups	38.4%	26.8%	17.8%	11.9%	5.1%	100.0%
		% within Quintiles of wealth score	20.3%	14.2%	9.3%	6.3%	2.7%	10.6%
		% of Total	4.1%	2.8%	1.9%	1.3%	0.5%	10.6%
	Acceptable Food Consumption	% within Food consumption groups	16.6%	19.2%	20.6%	21.4%	22.2%	100.0%
		% within Quintiles of wealth score	72.6%	84.4%	89.8%	93.3%	96.5%	87.3%
		% of Total	14.5%	16.8%	18.0%	18.7%	19.4%	87.3%
Total		% within Food consumption groups	20.0%	19.9%	20.1%	20.0%	20.0%	100.0%
		% within Quintiles of wealth score	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	20.0%	19.9%	20.1%	20.0%	20.0%	100.0%

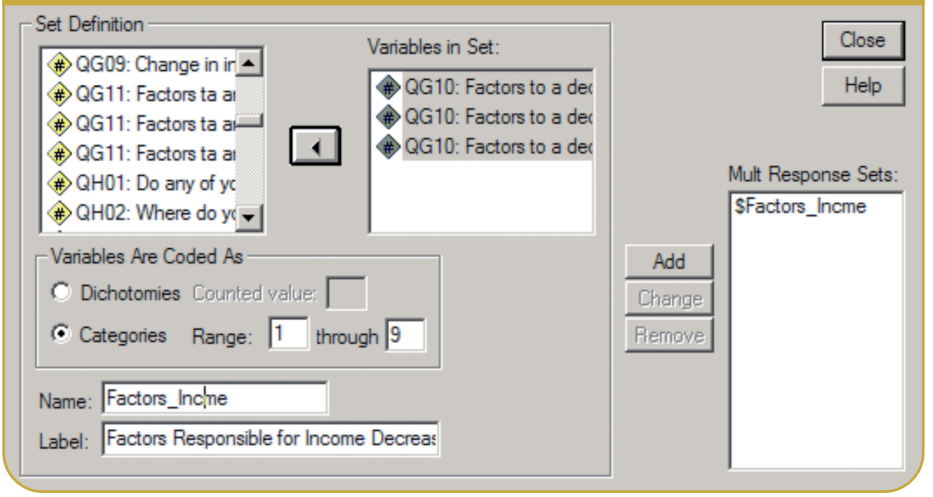
Table 4.23 shows the relationship between the food consumption of households and wealth status. It shows that food consumption increases stepwise by wealth quintile. More than 67 percent of households in the first wealth quintile reported poor food consumption, while 22.2 percent of households in the fifth wealth quintile and 21.4 percent of households in the fourth wealth quintile reported good food consumption.

4.4.3.4 Analysing multiple responses

A number of questions in CFSVAs call for multiple responses from the respondents. For example, a typical CFSVA questionnaire includes questions about major crops cultivated, main income sources, and food sources. All these questions generate multiple answers. A household might have obtained its food from different sources, including food bought from the market, produced, received as food for work, and borrowed from neighbours. It is a common practice to create separate variables for each of these answers, and these answers are mutually exclusive

The multiple response feature in SPSS allows one to analyse variables taking multiple responses into account. The first step is to define variable sets. All the variables containing (multiple) responses should be inserted into the “Variable in Set.” The next step is to select the type of variables included in the variable set and their range of values. After giving the name of the “Multiple Response Set,” add the variable set to the Multiple Response Set. Now the multi-response set is ready for analysis. To analyse, go to “Multiple Response,” then “Analyse,” and select the desired analysis.

Figure 4.12: Define multiple response sets



It is a good practice to paste the command to a syntax file for future reference. This will also help the analyst if she/he wants to regenerate the table.

Box 4.18 presents syntax to compute percentages. It essentially computes percentages of responses and percentages of cases.

Box 4.18: Syntax to compute percentages

```
mult response
groups=$Income_decline_factors 'Factors responsible for income decrease' (QG10_V1
QG10_V2 QG10_V3 (1,9))
/frequencies=$Income_decline_factors.
```

The output of the analysis is presented in Table 4.24. The column that presents percentage of responses calculates percentage of total responses. For example, in this case 623 households responded to the question "What are the factors that led to a decrease in your household income?" A household could answer 3 different responses from a list of 9. Sixty-five households identified 3 factors, 194 households identified 2 factors, and 364 households identified only 1 factor responsible for income decrease.

Altogether, 623 households responded to this question with 947 responses. The percentage of responses column calculates percentage of responses and the percentage of cases column calculates the percentage of households that responded to this question (623 in this example).

Table 4.24: Output of the analysis

		Responses		Total
		N	Percent	
Factors responsible for income decrease	Loss of employment	124	13.1%	19.9%
	Loss of crop/animal	65	6.9%	10.4%
	Prolonged illness of income earner	152	16.1%	24.4%
	Death of income earner	48	5.1%	7.7%
	Decrease in remittance income	8	0.8%	1.3%
	Loss of asset	147	15.5%	23.6%
	Exposure to natural disaster	124	13.1%	19.9%
	Market failure	157	16.6%	25.2%
	Other	122	12.9%	19.6%
Total		**947**	**100.0%**	**152.0%**

4.4.3.5 Measures of variation

Variance and standard deviation

Variance is a measure of dispersion around the mean of a continuous variable, equal to the sum of squared deviations from the mean divided by one less than the number of cases (degrees of freedom). The variance is, therefore, the average squared distance between the mean and the observations made (and so is a measure of how well the model fits the actual data). However, the variance is measured in units that are the square of those of the variable itself. The **standard deviation** is obtained by calculating the square root of the variance.

The **standard deviation** of a distribution, a measure of dispersion based on a deviation from the mean (which are squared, summed, and averaged and then the square root is taken), has the same unit as the original observations and can be used for data measured on an interval or ratio scale. A large standard deviation (relative to the mean; also called coefficient of variation) indicates that the data points are distant from the mean. In this case, the mean may not be an accurate representation of the data. A standard deviation of 0 would mean that all the scores were the same.

In a normal distribution, 68.27 percent of cases fall within one standard deviation on either side of the mean, 95.45 percent of cases fall within two standard deviations, and 99.73 percent fall within three standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95 percent of the cases would be between 25 and 65 in a normal distribution.

Confidence intervals

Confidence intervals enable analysts to make statements on the precision of their estimates. For example, if 30 percent of households were observed in the representative sample to be female headed, confidence intervals could be added to state that "the analysts are 95 percent sure that between 25 and 37 percent of households (in their sampling universe) are female headed." Confidence intervals should always be used in CFSVAs when reporting highly standardized indicators such as stunting, wasting,

underweight, low BMI, and low MUAC. Confidence intervals can be used in CFSVAs when reporting key indicators such as percentage of food insecure, percentage of poor food consumption. Confidence intervals are not typically reported for descriptive indicators in the text of a CFSVA; however, one should strive to include them in annex tables.[67]

4.4.3.6 Tests of significance

Significance is a statistical term that indicates how sure the researcher is that a difference or relationship exists. Tests of significance help the researcher and the audience know if differences between groups are real or by chance. When a statistic is significant, it simply means that one can be very sure that it is reliable and can be referred to the entire population. It does not mean the finding is important or that it has any decision-making utility. This significance, produced by the statistical tests discussed here, is referred to as the p-value (probability value).

The p-value can be interpreted as the probability of a difference occurring by chance alone. If all other biases are eliminated or accounted for, then one can assume that when this p-value is small, the differences are due to a factor other than chance. The cut-off for significance most often used is 0.05. If a p-value is less than 0.05, then assume that the relationship observed is real not by chance. Usually p-values are reported by their actual value to three decimal places, or as >0.05, <0.05, <0.01, or <0.001. Significance levels, when appropriate, are usually reported in the body of the report and in the annex tables.

In this section, some of the more commonly used statistical tests are presented. However, there is a wealth of further statistical tests, many available in SPSS. For a more complete guide to tests of significance, see Discovering Statistics Using SPSS,[68] or any other statistical manual or textbook.

It is very important to note that CFSVAs employ cluster sampling methods, which require special analytical approaches to calculating significance levels and confidence intervals. Standard packages such as the basic SPSS package, do not compute accurate p-values for surveys that are sampled using a cluster design. The appropriate statistical analyses can be obtained using the SPSS Complex Samples module or other special software.

Table 4.25 provides some guidance on what tests of significance to use when comparing different types of data. Keep in mind that this is a generic list and should be used only as a guide. It is the analyst who should decide which test is appropriate for what analysis based on a number of factors.

67. Automatic applications exist for computing confidence intervals for percentages. It is good practice to report them, at least for key indicators.
68. A. Field, 2005, *Discovering Statistics Using SPSS*, 2nd ed., London: SAGE Publications Ltd.

Table 4.25: Example of test of significance for different types of variables

	Dependent variable	Independent variable	When to use	Example	Procedure
Independent T-test	Continuous	Categorical binomial	To compare differences in the means of two groups (identified by the categories of the binomial variable) To see if the difference is statistically significant ($p<0.05$)	Compare the mean z-scores of male and female children	Run the independent samples T-test; Report the two means; Check if the T value is statistically significant ($p<0.05$)
One-way ANOVA: Post-hoc Multiple Comparisons	Continuous	Categorical	To compare differences in the means of three or more groups (identified by the categories of the categorical variable)	Compare the mean z-score by residence status (IDP, refugee, or resident HHs)	Run the One-Way ANOVA post-hoc procedure Check if the categorical variable explains in a significant way some of the observed variation through the F-test. Check which differences are statistically significant ($p<0.05$) through the post-hoc tests (e.g., REGWQ, Tukey HSD, Games-Howell, etc.)
Chi-square	Categorical	Categorical	To detect whether there is a statistically significant association between two categorical variables	Explore the association between food consumption groups and ethnic groups	Compute the Chi-square and report the value Check if the value is statistically significant ($p<0.05$) (The Chi-square helps determine whether the association is statistically significant)
Bivariate Correlation	Continuous	Continuous	To assess the general association between two variables (i.e.,one variable increases/decreases when another increases/decreases)	Correlation between children's height and weight	Compute the Pearson Correlation Coefficient and report the value Check if the correlation is statistically significant (two tailed tests) ($p<0.05$)
Simple Linear Regression	Continuous	Continuous/ Categorical binomial (0/1 values)	To measure how the dependent variable changes with a one-unit increase in the independent variable	Regressing food consumption score by wealth index	Run Simple Linear Regression Model Report R^2 adjusted, B value Check and report if B is statistically significant ($p<0.05$)
Multiple Linear Regression	Continuous	Two or more continuous/ categorical binomial (0/1 values)	To measure how the dependent variable changes with a one-unit increase in the independent variable (controlling by the other variables in the model)	Regressing food consumption score by wealth index and gender of the HH head	Run Multiple Linear Regression Model Report R^2 adjusted, B values Check and report if B values are statistically significant ($p<0.05$)
Multivariate General Linear Model (GLM)	Continuous	2 or more continuous variable and/or 2 or more categorical variables	GLM combines ANOVA and Regression to analyse the effects of more than one independent variable on the dependent variable (and to see how these independent variables interact)	Analyse the effects of ethnic group, province, and wealth index on the food consumption score	Run a Multivariate GLM Interpret the output from main ANOVA table Report R^2 adjusted, B values
Logistic Regression	Categorical	Two or more continuous variables and/or two or more categorical variables	To predict the probability of an event occurring for a given household	Predict which households are more likely to be food insecure according to the province of residence and WI	Run a Logistic Regression Check the overall fit of the model Check which variables significantly predict the outcome (in SPSS, check table "Variables in the equation")

In a typical CFSVA, the most commonly used tests of significance include the Chi-square test, z-test, t-test, and the ANOVA. For further information on tests of significance, consult a statistics manual.

4.4.3.7 Multivariate analysis

Multivariate analysis in statistics describes a collection of procedures involving analysis of more than one statistical variable at a time. In design and analysis, these techniques are used to perform studies across multiple dimensions while taking into account the effects of all variables.

Regression

Regression analysis is a technique used for the modelling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables (explanatory variables). The dependent variable in the regression equation is modelled as a function of the independent variables, corresponding parameters ("constants"), and an error term. The error term is treated as a random variable. It represents unexplained variation in the dependent variable. The parameters are estimated so as to give a "best fit" of the data. Most commonly the best fit is evaluated by using the "least squares method," but other criteria have also been used.

Data modelling can be used without there being any knowledge about the underlying processes that have generated the data; see http://en.wikipedia.org/wiki/Regression_analysis—cite_note-Berk-0#citenote-Berk-0; in this case the model is an empirical model. Moreover, in modelling, knowledge of the probability distribution of the errors is not required. Regression analysis requires assumptions to be made regarding probability distribution of the errors. Statistical tests are made on the basis of these assumptions. In regression analysis, the term model embraces both the function used to model the data and the assumptions concerning probability distributions.

Regression can be used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modelling of causal relationships. These uses of regression rely heavily on the underlying assumptions being satisfied.

Regression analysis is complex, and therefore cannot be adequately covered in these guidelines. However, a few key concepts and guiding principles are presented here. CFSVAs often, but not always, use regression analysis. Nutritional analysis makes frequent use of regression techniques.

a) When is regression used for CFSVAs, and why?

In CFSVA, regression is primarily used to understand causal relationships between the variables that are important for decision-making purposes. For example, to explore the relationships between stunting (height-for-age, an indicator of chronic malnutrition) and dietary quality while controlling for sanitation facilities, access to potable water, mother education, and household income, the analyst may want to set up an OLS (Ordinary Least Square) model, where height-for-age is a dependent variable and all of the other variables mentioned could be explanatory variables.

However, before estimation of the model, the analyst has to test for:

- **multicollinearity:** one or a combination of explanatory variables is strongly correlated to another explanatory variable;
- **heteroskedasticity:** the variance of the error terms changes across observations;
- **specification error:** the model is wrong, by missing important explanatory variables or by having other incorrect assumptions; and
- **endogeneity:** when an explanatory variable is itself a function of the dependent variable.

Necessary correctional measures need to be taken if the tests identify multicollinear variables, heteroskedasticity or omitted variables. If any of the explanatory variables is found to be endogenous, the variable has to be replaced by suitable instrumental variables. It is important to understand that a simple regression involving only the dependant variable and one independent variable is similar to a Pearson correlation.

b) Why control for other factors – confounders?

Confounders refer to factors that relate both to the dependent (outcome) and independent variable of interest. For example, we can hypothesize that children under 5 tend to be more underweight (low weight for age z-score) in female-headed households. We can run a simple compare means, or a t-test, to explore the differences in mean z-score between male- and female-headed households. However, the critical question is whether the sex of the household is the only factor responsible for the nutritional status of the children in the household? In a regression analysis, one could enter both education level and sex of the head of the household. It may be found in this analysis that the head of household no longer has a significant effect on underweight z-score, but that education level does. This could likely arise because female heads of household, in this example, often have lower education levels than male heads of households. The regression analysis controls for the difference in sex when it estimates the effect of education, and vice versa. Hence we can say "controlling for education, sex of household head is not significantly related to underweight z-score."

Female-headed households is still important as an identifier of vulnerable households for targeting, since gender may be more easily identified than education level, even if we know that the real reason is that most female household heads have a low education level.

c) Why explore interactions?

Interactions illuminate how a cause of food insecurity might be modified by another variable. For example, if we look at sanitation, water source, and underweight status of children under 5, we might find that improving sanitation has no effect unless in the presence of a safe source of water. In the regression, the two variables (sanitation and water) are simply multiplied together to give the interaction term. Environmental variables, economic factors, education, and age are common effect modifiers (variables that result in statistical interactions).

d) Coefficient of determination

In linear regression models, R^2 is a statistic that gives some information about the merit or fit of a model. R^2 is the square of the correlation coefficient between the dependent variable and the estimate it produced by the independent variables, or equivalently

defined as the ratio of regression variance to total variance. It is a measure of determination of how well the regression line approximates the real data points. An R-squared of 1.0 indicates that the regression line perfectly fits the data, and 0.0 indicates that one term does not help to know the other term at all. For regression applied to household surveys, it is normal to find an R^2 between 0.15 and 0.25, or (exceptionally) a little higher.

Principal component analysis and cluster analysis[69]

This section discusses the following two multivariate analysis techniques:

- Principal component analysis (PCA) – which belongs to the factor analysis family; and
- Cluster analysis – which belongs to the classification family.

Both techniques can typically be used to reduce the complexity of the data set for exploratory purposes: factor analysis, to reduce the selected variables into a lesser number of factors; and cluster analysis, to group all cases into a number of groups. Detailed explanations of how PCA and cluster analysis techniques work are beyond the scope of these guidelines; interested readers should refer to specialized textbooks for more information. This section presents a simplified summary of the statistical ideas on which PCA and cluster analyses are currently used in CFSVA and FSMS, as well as of the terminology used throughout those analyses.

In addition to SPSS, several commercial statistical software packages perform both PCA and cluster analysis. With the support of WFP and FAO, VAM typically also uses a software developed explicitly for socio-economic and food security analysis (ADDATI, or the brand new Windows version ADDAWIN). This software was designed for the use of food security specialists. It includes preselected algorithms proven to be suitable when analysing socio-economic and nutrition data for food security and vulnerability analyses. It uses the output of the PCA as the input for the cluster analysis and facilitates the final interpretation of the outputs providing the cluster results in terms of the original input variables.

ADDATI/ADDAWIN cannot perform factor analysis with rotation. For this type of multivariate analysis, the software normally used in VAM is SPSS.

Principal component analysis

Factor analysis is used to study the patterns of relationships among many dependent variables, with the goal of discovering the underlying variations that affect them. The inferred underlying variables are called factors. Principal component analysis (PCA) uses a factor extraction method to form uncorrelated linear combinations of the observed variables. The first component explains maximum variance. Successive components explain progressively smaller portions of the variance and are all uncorrelated with each other.

PCA is one technique of multivariate analysis that applies to continuous variables. The objective of PCA is twofold:

69. This chapter strongly benefits from inputs and quotations from: WFP/VAM, *Household Food Security Profiles (Thematic Guidelines)*, April 2005; S. Griguolo, *ADDATI Users' Manual*, July 2003, IUAV; S. Landau and B.S. Everitt, *A Handbook of Statistical Analyses Using SPSS*, 2004; Chapman & Hall/CRC, Andy Field, Discovering Statistics using SPSS, 2005; SAGE ADDATI help; and SPSS help.

- to identify and describe the underlying relationships among the variables by creating new indicators (called "factors" or "principal components") that capture the essence of the associations between variables; and
- to reduce the complexity of the data, saving a limited number of these new variables that is sufficient to keep the most relevant aspects of the description with a minimal loss of detail.

PCA yields as many principal components as there are initial variables. However, the contribution of each principal component to explaining the total variance found among all variables will progressively decrease from the first principal component to the last. As a result, a limited set of principal components explains the majority of the matrix variability, and principal components with little explanatory power can be removed from the analysis. The result is data reduction with relatively little loss of information.

It is recommended to use the rotated solution in most circumstances. Rotation of the result will give a new solution: the new factors explain the same variance and can be much better interpreted as the underlying dimensions. The analyst can understand the "meaning" of each dimension and then decide which of the uncovered dimensions he wants to use in subsequent analyses.

Uses of PCA

Two of the most common uses of PCA in CFSVA analysis are briefly described here:

Scoring on the first principal component (single factor solution)

PCA might be used to build a synthetic composite index moving from more than one variable. In this case, the first principal component is taken as the new variable on which statistical units might be measured and ranked. For example, variables from a household survey that are associated with wealth (quality of housing, assets owned by the household, etc.) can be used to perform a PCA. Since what those variables have in common is related to wealth, the first component can be interpreted as a **wealth index**.

While the single factor solution is very straightforward and easy to interpret, its use should be limited to specific cases. If the variables are uni-dimensional, then an exploratory analysis showing a single factor grouping, where all items "hang together," is supportive. If the variables used in the analysis have multiple facets or dimensions, a single factor solution would not be able to maintain a minimum description for all of the analysed statistical units. In this case, trying to capture the underlying variations of all of the input variables through a unique index would undermine the instrument's construct validity. That is because the use of a small selection of measuring items (i.e. the information maintained by one factor only) could lead to false and confusing results that would not reflect the complexity of the original data.

Using principal component(s) as input variable(s) for follow-up analysis (specifically cluster analysis)

PCA creates a set of new variables or components that, being perfectly uncorrelated, explain different portions of the original total variance.

As one of the main purposes of PCA is to reduce the dimensionality of the data set, components are ranked by their decreasing contribution to explaining the total

variance. It is hence possible to remove components with little explanatory power. If the main purpose of PCA is to describe statistical units on the basis of the relationships among selected variables, data reduction is a secondary objective. Furthermore, if the final aim is to cluster units based on those relationships, it is recommended that analysts keep as many principal components needed to capture up to 80 percent of the total variance. Such a high level of consistency with the original complexity of the data would ensure a good reflection of the relationships among variables. It would also guarantee that particular combinations of variables' values were maintained and not smoothed too much through a high data reduction approach.

When data reduction is the primary objective, the analyst may want to remove more components. One rule of thumb is that the Eigen value of each extracted component should be higher than 1; an alternative is to keep as many factors (after rotation) that still have a readily understandable meaning. For a subsequent clustering, the analyst can even exclude factors irrelevant for the clustering activity to be undertaken.
For example, average decadal rainfall, Normalized Difference Vegetation Index (NDVI), and other climatic data were used to conduct a factor analysis in Sudan. Four underlying, meaningful factors were retained. The third factor, related to rainfall during the dry season, was not used for clustering, since rain during that season has little economic importance.

Cluster analysis

Clustering data is a common technique for statistical data analysis. Clustering is the classification of similar objects into groups or, more precisely, the partition of a data set into subsets (clusters) so that the data in each subset share some common features, often proximity according to some defined distance measure.

Note that cluster analysis is an exploratory data analysis tool that aims at sorting different objects into groups such that the degree of association between two objects is maximal if they belong to the same group, and minimal otherwise.

Clustering is one of the key parts of many large data set analyses. In fact, it is not possible to analyse and describe the situation of each statistical unit (be it a household, a village, a district, a region, etc.) separately, since there might be too many. There is a clear need to identify main patterns of similar characteristics. Clustering involves some kind of subjectivity based on analysts' choices of specific methods for clustering.

In conducting cluster analysis, analysts are often faced with two questions: How many clusters are there in the data set? and What is the compactness (inertia) of each cluster?

How many clusters in the data set?

A given data set does not contain a definitive number of clusters. First, because cluster analysis involves a series of iterations performed by statistical software, there will be some variance in the number of clusters and the assignment of particular households to clusters each time the analysis is performed, depending on the initial "cluster seeds." Second, several different methods and algorithms can be used to produce clusters, and the number of clusters produced will vary depending on the type of clustering method used. For very large data sets, the partition method, with a random selection of the initial centres, seems to be most appropriate. Specific algorithms to

improve the quality of a partition are implemented, being different in different software packages.

What is the compactness of each cluster (inertia)?

The measurement of the dispersion or compactness of each cluster is called **inertia or internal variance of the cluster.** The degree of inertia within and among clusters provides a useful means of determining the final number of clusters that best fits the data.

There are no standard thresholds indicating what level of inertia is good, acceptable, or poor, and the final decision remains with the analyst. However, the ratio between the inertia of the overall cloud (the dispersion found among all units in the dataset) and the inertia within each cluster should be maximized. Doing so ensures that the similarity among units belonging to the same cluster (e.g. within clusters) is high, while the similarity between clusters is very low (e.g. maximizing intra-cluster homogeneity and inter-cluster heterogeneity).

One of the strengths of ADDATI (or ADDAWIN) is that it incorporates a specific formula (objective function) to calculate the intra- and inter-cluster inertia as a measurement of partition optimality with a given number of clusters. In other words, it measures how compact a set of clusters is.

In addition, the clustering in ADDATI displays a graph that plots how the value of the objective function decreases when the number of clusters retained is increased. By inspecting this graph, the user can focus on one or more promising partitions, with a number of clusters within the range he/she would like to obtain and a value of the objective function sufficiently high. This tool indeed helps the analyst find a number of final clusters, which is a fair **trade-off** between the level of synthesis achievable (few clusters are always more convenient) and a significant level of homogeneity of characteristics within the clusters (provided by the value of the objective function that represents the rate of information maintained).

One of the common uses of cluster analysis in CFSVA is to create groups of households with similar food consumption patterns, for profiling purposes or for further analysis. Typically the principal components are used to create the clusters.

Cluster analysis is also used to categorize households that share similar livelihood strategies into **livelihood groups**. The aim of CFSVA livelihood grouping is not to replace a comprehensive livelihood analysis but to utilize livelihood strategies as a basis for classifying populations.

4.4.4 Key references: Household data analysis

- Chapman and Hall/CRC Andy Field. 2005. *Discovery Statistics Using SPSS*. Sage ADDATI help; and SPSS help.
- Griguolo, S. 2003. ADDATI User's Manual. IUAV, July.
- High, R. 2000. *Dealing with Outliers: How to Maintain Your Data's Integrity.* Computing News. UO Computing Center. University of Oregon.
- Landau, S., and B. S. Everitt. 2004. *A Handbook of Statistical Analyses Using SPSS*, http://cc.uoregon.edu/cnews/spring2000/outliers.html.
- Shawna, J., K. Marcus, C. McDonald, T. Wehner, and M. Palmquist. 2005. *Introduction to Statistics.* Writing@CSU. Colorado State University Department of English. Retrieved 12/31/2007 from http://writing.colostate.edu/guides/research/stats/.
- United Nations, Department of Economic and Social Affairs, Statistics Division. 2005. *Household Sample Surveys in Developing and Transition Countries.* Studies in Methods, Series F No. 96.