

ANNEX 2

Guidelines on sampling design and procedure for surveys of pesticide use and hazard awareness in rural communities

ANNEX 2A – Survey sampling procedure (english)

ANNEX 2B – Survey sampling procedure (french)

ANNEX 2C – Guidelines for Key Indicators Surveys

ANNEX 2A

Comments on the use of the Questionnaire for survey of rural community use of pesticides and awareness of pesticide hazards, and suggestions for survey sampling procedure

Colin Tingle, Ian Grant and Ken Campbell

The following is a basic guide to the key steps in designing the sampling procedure for use of the questionnaire within the target communities. For further detail, see the Guideline for sampling below - taken from the Key Indicators Survey (KIS) Guidance Manual <http://www.measuredhs.com/aboutsurveys/kis.cfm> (Annex 2C)

See other information about surveys also at <http://www.measuredhs.com/aboutsurveys/>

Sampling Procedures

Assuming that the *target population* is mainly going to be of small farmers or of households that may be involved in farming as well as other activities, the *sample unit* is therefore based on households.

The plan is to do a cross-sectional survey in the area and the number of households to be included in the study will be determined using a sample of all households in the area. This was well laid out by the participants from Ethiopia, and is recommended - so the following is based on their summary of the procedure for calculating a sample size.

$$n = (Z\alpha/2)^2 p (1-p)/d^2$$

Where

- n = Sample size
- (Z $\alpha/2$) = Reliability coefficient = 1.96
- p = proportion using pesticides - assumed to be 50%.
- d = assumed marginal error (5%)
- $n = (1.96)^2 (0.50) (0.50)/(0.05)^2 = 384$
- 10% (or greater) non-response rate should be added to the final sample size. Accordingly,
- $n = 384 + 38 = 422$

So, plan for a sample size from all villages combined of about 450, this should be a minimum that is adequate for the current purposes. Any more than this will, of course, improve the data (see below).

A first step in actually designing the survey procedure will be to obtain a reasonably good list of households (e.g. from national databases or other information sources or from the village leaders).

The next step is to divide the samples appropriately between wards and villages, or other appropriate administrative units.

As an example, the National Census data in Tanzania for the wards mentioned in the mini-project outline is as follows (Table 1), and this Table gives an indication of the how the sample might be spread between the wards and villages.

Table 1. Data on Mbeya District and wards, from Tanzania National Census 2002 (<http://www.nbs.go.tz/>)

Reg_id	12	12	12	12
Region	Mbeya	Mbeya	Mbeya	Mbeya
Dist_id	1207	1202	1204	1208
District	Mbarali	Mbeya Rural	Rungwe	Mbeya Urban
Ward_id	11	3	29	8
Ward_shehi	11 Igurusi	3 Tembela	29 Kiwira	8 Igawilo
Type	Mixed	Mixed	Mixed	Mixed
Male	11217	6627	9605	4835
Female	12251	8079	10514	5669
Total	23468	14706	20119	10504
Household numbers	6421	3836	4920	2840
Average household size	3.7	3.8	4.1	3.7
Area_km ²	214.13	87.11	115.35	7.18
Density	109.6	168.82	174.42	1462.95

The total households in all these wards is 18,017. By splitting the sample according to the relative number of households per ward and finally per village, you end up with:

Ward_shehi	Igurusi	Tembela	Kiwira	Igawilo
Percent of households	35.6%	21.3%	27.3%	15.8%
Number of villages in sample	3	3	2	1
For a total of 450 questionnaires	160	96	123	71
Samples per village (approx)	53	32	62	71

When it comes to doing the survey in a village, if there is a total number of households known for the individual village, divide the total by the sample number for the village (e.g. 71 households to be sampled in the village within Igawilo Ward, gives $422/71 = 5.94 \cong 6$), and

then take this figure as the interval between households. However, in general it is not recommended to exceed an interval of every tenth household as this becomes tiresome for the data collectors.

Start at a randomly chosen household in order to avoid bias. If possible make sure that the sample covers both the central parts of a village, as well as the households towards the outside of the village. If there is a road running through the village, make sure that selected households are not only those that lie along the road, but that more remote households are also included in the sample. One way of doing this is to sample along approximate North-South, and East-West lines starting in the centre of the village. This latter needs to be considered on a case by case basis as each village will have its own characteristics and it is difficult to lay down specific rules.

NB. The decision on sample size is not related to the actual number of households present. However, it should be noted that with the example given above for the Tanzanian mini-project, a sample size of approximately 450 households (already quite taxing, based on the time and the resources available to the mini-projects) represents only about 2.5% of the total households in the target wards. It is vital that this is taken into account in the interpretation of the results with regard to their representativeness to the whole target region. From a statistical perspective, it is a very small sample to be truly representative of the area as a whole!

Guidance on use of the questionnaire

Open Questions:

When using the questionnaire in the villages, the best procedure is to ask ALL questions as open questions. In other words do not prompt the respondent by giving a choice of answers – until they have given an answer. The enumerator can then tick the relevant box. If the answer they provide is not clear, then it is possible to come back to them with more explanation and the choices, but stress that they can (in most cases) provide an alternative answer under the “other” category.

As an example, in part 3 of the questionnaire, for the section on reporting of incidents, ask the question “To whom would you report any incidents to?” but don’t give them any of the listed choices. See what they say. If they don’t know who to report an incident to then you can enter “don’t know” under other (specify). You can also tick more than one box – they may for example say that they would report it to the village environmental committee and to the school teacher – so tick the box for one and write the other against “other”.

The enumerators will need to have a few copies of the introductory page to the questionnaire (see Annex 1) or a similar sheet available to provide the background of the survey to the respondents. The same sheet can also have guidance notes for enumerators on it. As long as the name of the questioner and individual number of the questionnaire are on each sheet, individual questionnaires can be cross referenced against a set of master sheets with the remaining details. Page one is therefore kept as a separate file at the beginning of Annex 1.

It is assumed that the survey will be a team effort and will be using a number of assistants/enumerators to conduct the surveys. Accordingly, the assistants will need to be familiar with the questionnaire and the procedures to be followed. It is important that all assistants receive the same briefing so that data collection is consistent and that individual approaches do not introduce an element of bias into the survey results. A short test run is advised – so make a few copies of the questionnaire for this first. This can assure consistency of approach to questioning by enumerators and also highlight any lack of clarity in the questions themselves, so allowing the test run to lead to changes that improve the questionnaire, if necessary.

When it comes to asking about the names of different chemicals/products involved these may be known by a variety of different names – local as well as brand names. The actual chemical may not always be stated on the packaging. It is also possible that some chemicals are in containers without any labels at all. Therefore make sure that as much information as possible is collected here to assist in determining what the pesticides actually are. .

If possible, either as part of this survey or as part of a follow-up exercise, one can ask questions at the places where the people purchase the chemicals. When asking questions of possible suppliers of the various chemical products involved, it is hoped that there will be some information on where the households actually purchase them coming from the household questionnaires. Follow-up surveys of the vendors should then be possible relatively easily. The information required from vendors is simple. Basically: What do they sell, where do they obtain it, what sort of containers is it sold in and does it come with instructions? Are there brands or types of chemical any that are more popular than others – they may be able to say how much they sell. In addition, one would need to have some information on where and how it is stored by the vendor - e.g. do they follow any guidelines for storage? Have they been given any guidelines for storage? Are they aware that the chemicals can be hazardous?

A WHO form, that may be of interest for any follow-up surveys where health-centres are involved, and guidance documents to accompany it can be accessed at:
<http://www.wpro.who.int/hse/pages/exposure.html> (click on record form or guidance documents).

If any health-related data might be collected in the future, the work already done by WHO in standardising these types of records should as far as possible be followed.

ANNEX 2B.

Commentaires sur le questionnaire Ecotox préparé par ODI Sahel devant servir à l'enquête écotoxicologique à Mopti, au Mali, sur la taille de l'échantillon et la procédure d'échantillonnage

(Suggestions de Colin Tingle, Ian Grant et Ken Campbell)

Contexte

Il est nécessaire d'avoir quelques informations sur les ménages. C'est dans la partie 1 qui nous amène à poser la question de savoir à quel type de cultures et à quel type d'élevage les familles s'adonnent-elles ? Après cela, nous chercherons à savoir si oui ou non ces ménages utilisent des pesticides. Cependant, cela doit se faire en tenant compte non seulement du fait que les villageois peuvent potentiellement utiliser d'autres intrants mais également des investissements totaux des villageois ou du niveau de dépendance à l'agriculture/pêche. Prenant en compte les deux extrêmes qui se présentent à nous, si par exemple le ménage n'a qu'un petit poulailler et n'adopte qu'une agriculture de subsistance, les capacités financières d'acquérir des pesticides sont limitées ou inexistantes. Si le ménage adopte un système agricole intensif, c'est-à-dire avec des surfaces importantes de cultures de rente et/ ou un élevage de vaches laitières par exemple, donc leur productivité est sûrement dépendante d'une utilisation intensive d'intrants, y compris les pesticides. Leur façon d'utiliser les outils manuels/ bœuf ou âne ou encore les pirogues et équipements de pêche ainsi que leur contraintes observées dans la façon de gérer leur bétail sont également des informations qui entrent dans le même cadre.

Tranches d'âge:

Pour ce qui est de l'âge de l'interviewé, il sera tout juste nécessaire de relever son véritable âge. Comme vous l'avez suggéré, Il serait bien de demander l'âge des membres de la famille et le système d'enregistrement le plus facile qui soit consiste à les classer par tranche d'âge. Cependant, il faudra vous assurer que ces tranches d'âge sont compatibles avec celles utilisées dans les publications/rapports nationaux de recensement. Deux groupes sont utilisés dans les publications/ rapports nationaux de recensement.

- Les tranches d'âge: 0-14, 15-64, 65 et plus.
- La tranche d'âge 15-64 a été dans certains rapports modifiée comme suit: 15-29, 30-44, 45-64,
- dans certaines publications de recensement
- Les tranches d'âge avec un décalage de 5 ans :
0-4 , 5-9 , 10-14 , 15-19 , 20-24 , 25-29 , 30-34 , 35-39 , 40-44 , 45-49 , 50-54 , 55-59 , 60-64
65-69 , 70-74 , 75-79 , 80+

Pour cela, il est donc nécessaire d'utiliser des données qui soient compatibles avec celles du rapport de recensement. Les tranches d'âge les plus importantes sont les suivantes : 0-14, 15-29, 30-44, 45-64 et 65 et plus. Si cela s'avère difficile pour les enquêteurs, ils n'auront juste qu'à enregistrer le nombre de personnes dans la famille.

Cultures

Je ne suis pas en mesure de vous dire quelles seront les cultures les plus prisées des villageois. Une liste de culture est suggérée dans la partie 1 du questionnaire. Veuillez revoir

cette liste et la modifier au besoin et suivant les informations que vous possédez. Etant donné le très grand nombre de possibilités, il est suggéré d'inclure les cultures les plus fréquentes et ensuite de laisser deux ou trois rangées pour « autres ». L'idée est d'obtenir une impression générale sur ce qu'ils cultivent, c'est-à-dire de savoir si c'est des cultures de subsistance ou de rente et si possible savoir quelle est la culture la plus importante (ou quelles sont les deux cultures les plus importantes). Si cela s'avère être des cultures de rente et que par ailleurs les habitants cultivent du coton par exemple, il serait possible qu'ils a) soient familiers à et b) qu'ils utilisent quelques ou tous les intrants chimiques dont l'usage est recommandé sur ces cultures. Si par contre, les habitants ne cultivent que du mil, du sorgho ou encore des légumes, donc, les intrants utilisés ne sont sûrement pas les mêmes. (même si les légumes en tant que culture de rente peuvent requérir une quantité importante d'intrants chimiques). Biologiques ?

Les produits biologiques constituent un autre point que vous pourriez prendre en considération- par exemple, le café biologique ? Existe-t-il des probabilités dans la zone qui fait l'objet de l'enquête ? D'après certaines informations que j'ai trouvées sur Internet, il existe des accords de réciprocité dans les transactions commerciales internationales pour ce qui concerne le café biologique dans la localité de Mbeya qui est commercialisé au Canada sous l'appellation Café Mbeya. Au besoin, vous pourriez répondre par un oui/non dans une colonne que j'ai ajoutée dans la liste des cultures si jamais la culture biologique se pratiquait dans le village. Cependant, vous êtes libre de ne rien mettre ou tout simplement de rayer cette colonne du questionnaire si vous jugez qu'elle n'est pas nécessaire dans le cadre de l'enquête.

La pêche

Nous vous proposons d'élaborer une série de questions qui seront destinées aux pêcheurs de la zone afin de fournir une base de données similaire sur leurs moyens d'existence (par exemple, savoir si c'est une pêche de subsistance ou destinée au marché local).

Le bétail

Il serait précieux de collecter des informations sur l'utilisation de certains produits chimiques par les éleveurs. Par exemple, le nombre de ménages qui utilisent l'épandage/ bain parasiticide/ application/ autre/aucun contrôle sur les tiques (et autres statistiques sur les déprédateurs du bétail).

Usage des pesticides

A large spectre/ déprédateur spécifique : il se pourrait que vous soyez tenu de vous assurer que ces termes sont bien compris. Cette question est cependant facultative et pourrait être supprimée du questionnaire si vous le désirez.

Questions ouvertes:

Nous pensons que la meilleure approche pour mener à bien notre enquête et tirer un maximum d'informations est de poser des questions ouvertes. En d'autres termes, nous vous demandons de ne pas fournir aux interviewés une série de réponses avant qu'ils n'en fournissent eux-mêmes. L'enquêteur peut ainsi cocher la case appropriée. Si la réponse fournie n'est pas claire, il vous est possible de revenir à eux pour de plus amples explications et de choix, mais mettez l'accent sur le fait qu'ils peuvent (dans la majorité des cas) fournir une réponse alternative dans l'autre « catégorie ».

A titre d'exemple, dans la partie 3, notamment la section relative au report d'incidents, posez la question "A qui reportez vous les incidents qui se produisent?" mais ne leur fournissez aucun des choix figurant sur la liste pour ne pas les influencer et écoutez ce qu'ils vous diront. S'ils n'arrivent pas identifier la personne ou organe auquel ils rapportent les incidents, c'est-à-dire, s'ils n'ont personne à qui notifier ces incidents, vous pourriez donc mentionner « ne sait pas » dans autre (veuillez spécifier). Vous pourriez également cocher plus d'une case – ils peuvent par exemple dire qu'ils notifient ces incidents au comité environnemental du village et aux enseignants – donc cochez la case pour une réponse et mentionnez l'autre dans « autre ».

Procédures d'échantillonnage

Nous supposons que les échantillons concerneront principalement les petits exploitants/ les pêcheurs ou les ménages impliqués dans l'agriculture et autres activités. L'échantillon se fera sans doute au sein des ménages.

Les chefs de villages pourraient vous aider à avoir une liste des ménages. (D'autres peuvent également vous aider à l'avoir, vous êtes mieux placés pour en décider). L'objectif est de mener une enquête transversale dans la zone et le nombre de ménages qui sera pris en compte dans le cadre de l'enquête sera déterminé en utilisant un échantillon de tous les ménages de la zone. Cela a bien été spécifié par les participants issus de l'Ethiopie (Tadess Amara & Asferachew Abate) dans leur modèle d'enquête et c'est exactement ce que je recommande – ainsi, on s'est basé sur le résumé de la procédure utilisée pour calculer une taille d'échantillon. Voici ce que cela a donné :

- $n = (Z_{\alpha/2})^2 p (1-p)/d^2$
- où, $(Z_{\alpha/2}) =$ Coefficient de fiabilité = 1,96
- n = Taille de l'échantillon
- p = est suppose être de 50%.
- d = erreur marginale supposée (5%)
- $n = (1,96)^2 (0,50) (0,50)/(0,05)^2 = 384$
- 10% de taux de non- réponse doivent être ajoutés à la taille finale de l'échantillon. Donc, $n = 384 + 38 = 422$

Donc, calculez environ 450 échantillons si tous les villages doivent être pris en compte. C'est le minimum requis pour mener à bien l'enquête. Il est clair qu'un échantillonnage plus important vous permettrait d'avoir des données plus précises.

A supposer que le nombre total de ménages dans la zone soit de 18017 divisés en quatre districts suivant le pourcentage présenté ci- bas (tableau 1). En divisant l'échantillon suivant par le nombre relativement exact de ménages, vous obtiendrez les données suivantes :

Table 1

District	District 1	D 2	D 3	D 4
pourcentage des ménages	35.6%	21.3%	27.3%	15.8%

nombre de villages concernés par l'échantillonnage	3	3	2	1
pour un total de 450 questionnaires	160	96	123	71
échantillons par village amples per village (approx)	53	32	62	71

Pour ce qui est de mener l'enquête dans un village, si vous connaissez le nombre total de ménages dans un village donné, divisez le total par le nombre d'échantillons pour le village en question (par exemple, 71 ménages doivent faire l'objet d'un échantillonnage dans le village du district 4) et ensuite prenez ces chiffres ($18017 * 15,8/100/71=40$) comme intervalle entre les ménages. Commencez l'enquête au hasard pour éviter d'avoir des résultats biaisés. Cependant, il n'est généralement pas recommandé de dépasser un intervalle de 10 maisons étant donné que l'exercice pourrait par la suite devenir fatigant pour les enquêteurs. Ainsi, dans ce cas, sélectionnez chaque fois jusqu'à la 10^{ème} maison et non la 40^{ème}. Si possible, assurez-vous que l'échantillon couvre aussi bien le centre que les ménages qui se situent à l'entrée du village. Si le village est traversé par une route, assurez-vous que vous n'avez pas sélectionné que les ménages qui se situent le long de la route et que les ménages les plus en retrait ont également été inclus dans l'échantillon. Vous pourriez par exemple démarrer l'échantillonnage du Nord au Sud, de l'Est à l'Ouest du village en commençant par le centre. Cela doit cependant être traité au cas par cas étant donné que chaque village a ses spécificités et qu'il est difficile d'établir des règles communes.

Il serait utile d'avoir une page introductive sur laquelle il sera mentionné le titre de l'enquête ainsi que les informations sur les raisons de cette enquête, qui vous êtes, etc. chaque enquêteur devra avoir quelques copies ou une feuille pour briefer les villageois sur les pourtours de l'enquête. Cette feuille pourrait renfermer quelques directives qui pourraient également être mentionnées au verso. Tant que le nom de l'enquêteur et le numéro du questionnaire sont mentionnés sur chaque feuille, les questionnaires individuels peuvent être mentionnés parallèlement sur des feuilles avec les détails manquants. Je suppose que vous vous ferez assisté dans votre tâche par des personnes. En conséquence, ces assistants devront se familiariser aux questionnaires et aux procédures à suivre. un petit test pourrait s'imposer- donc, commencez par faire quelques copies de ce questionnaire à cet effet.

Concernant la question sur les noms des différents pesticides/produits chimiques utilisés, plusieurs autres appellations pourraient être utilisées par les villageois pour un seul produit notamment les appellations locales et le nom de commercialisation. Le vrai nom du produit n'est pas toujours mentionné sur l'emballage. Il est également possible qu'aucune étiquette ou notice ne soit inscrite sur le contenant des produits. Par conséquent, assurez-vous d'avoir obtenu le maximum d'informations pour pouvoir déterminer quels sont réellement les pesticides utilisés.

Si vous en avez l'opportunité, soit dans le cadre de cette enquête ou d'un exercice de suivi, utilisez le questionnaire 5 pour en savoir plus sur où est-ce que les villageois acquièrent ces pesticides. En posant ces questions aux éventuels utilisateurs de ces produits chimiques, on a espoir d'avoir des informations sur les véritables fournisseurs. Ainsi, les enquêtes de suivi qui seront menées avec les vendeurs seront relativement plus aisées.

ANNEX 2C

Reproduced from: <http://www.measuredhs.com/aboutsurveys/kis.cfm>

Users are encouraged to visit the Key Indicators Survey (KIS) website and examine relevant material online and download guidance documents. The KIS tool kit includes questionnaires, interviewer's manuals, and guidelines for sampling. The document on guidelines for sampling is reproduced below.

GUIDELINES FOR SAMPLING

This section presents guidelines for sampling for the Key Indicators Survey (KIS). The general principles that should guide the KIS survey sampling strategies are discussed first. The specific issues that should be considered in designing and selecting the sample for a KIS survey are then reviewed.

4.1 General Principles

Scientific sample surveys provide a relatively inexpensive and reliable way to collect social, demographic and health data on a large scale. In order to achieve consistency and high quality results, survey sampling activities should be guided by a number of general principles.

Survey coverage

The target population for the KIS survey depends on the type of survey. For the Family Planning, Maternal Health, and Infectious Disease surveys, the target population is all women age 15-49, while for the HIV survey, the target population is both women and men age 15-49.¹ For the Child Health survey, the population of interest is children under age five, with the survey respondent being either the child's mother, father or caretaker. Since all these target populations can be easily found in residential households, KIS is a household-based survey.

A KIS sample should cover 100 percent of the households in the desired survey area (e.g., project implementation area, selected districts or provinces, catchment area). In some cases, exclusion of some areas may be necessary because of extreme inaccessibility or insecurity; however, it is preferable if this can be considered at the beginning of the survey planning.

Probability sampling

Scientific probability sampling is the only way to achieve unbiased survey results. It also is the only methodology by which to estimate sampling error—the effect of interviewing a portion instead of the whole universe of interest. A probability sample is defined as one in which the units are selected randomly with known and non-zero probabilities. The term excludes purposive sampling, quota sampling, and other uncontrolled non-probability methods, since they cannot provide precision and/or confidence evaluation of survey findings.

¹ In some surveys, it may be desirable to restrict the target population to ever-married women or men.

Sampling frame

A probability sample can only be drawn from an existing sampling frame, that is, a complete list of statistical units covering the entire target population. The most commonly used sampling frame is a recent population census. Censuses usually provide good sampling frames because they utilize enumeration areas (EAs), which are small geographic areas of known population size. Nevertheless, an evaluation of the quality and the accessibility of the frame should be undertaken prior to sample selection. In the absence of an adequate, preexisting sampling frame, KIS survey managers should arrange to construct a list of villages or communities in the survey area with all necessary identification information and a measure of the size for each. A commonly used measure of size is the number of households residing in the village or community.

Simplicity of sampling design

In large-scale surveys, non-sampling errors are usually the most important sources of error and are expensive to control and difficult to evaluate. It is important to minimize this type of error in survey implementation. Therefore, the sample design for a KIS should be as simple and straightforward as possible to facilitate accurate implementation. ORC Macro's experience from the Demographic and Health Surveys (DHS) program shows that a two-stage sampling design is the most appropriate, as discussed later in these guidelines.

Care in sampling implementation

Care in sampling implementation is the last element in achieving good sampling precision. No matter how carefully a sample is designed and how complete the materials for conducting sampling activities are, if the implementation of the sampling activities by survey staff (office staff responsible for selecting sample units, field workers responsible for the mapping and household listing and interviewers responsible for correctly identifying and visiting the selected households) is not performed exactly as designed, serious bias and misleading results may occur.

4.2 Survey Domains

A *survey domain* is a subpopulation for which separate survey results are required. In designing a KIS sample, one of the first decisions to be made relates to the domains for which information is desired, e.g., are separate estimates of KIS indicators needed for urban and rural areas, for geographic or administrative units, or for individuals with different educational levels? The survey domains or *study or reporting domains* should be specified at the survey design stage so that decisions about the overall size of the sample can take into account the need to report results for the various domains. The total sample size generally represents the sum of sample sizes needed to provide estimates at a desired level of precision in all exclusive domains. Generally, the greater the number of domains, the larger the size sample size required.

4.3 Sampling Frame

A *sampling frame* is a complete list of sampling units that entirely covers the target population. The existence of a sampling frame allows a probability selection of sampling units. For a multi-stage survey, a sampling frame should exist for each stage of selection. The availability of a suitable sampling frame is a major determinant of the feasibility of conducting KIS survey. This issue should be addressed in the earliest planning for a survey.

A sampling frame for KIS could be an existing sampling frame, an existing master sample, or a sample of a previously executed survey of sufficiently large sample size, which allows for the selection of subsamples of desired size for the KIS.

The sampling frame used for KIS should be as up-to-date as possible. It should cover the whole survey area, without omission or overlap. Maps should exist for each area unit or at least for groups of units with clearly defined boundaries. Each area unit should have a unique identification code or a series of codes that, when combined, can serve as a unique identification code. Each unit should have at least one measure of size estimate (population and/or number of households). If other characteristics of the area units (e.g., socioeconomic level) exist, they should be evaluated and retained because they can be used for stratification.

Among the most common frames is a list of *enumeration areas* (EAs) from a recently completed population census. A typical EA is a small geographic area with clearly delineated boundaries in which 100-150 households reside. Before a census frame is used, it should be thoroughly evaluated; of particular importance is the existence of maps showing the boundaries of all of the EAs included in the frame.

A preexisting master sample (which is most typically a random sample from the census frame) may be used in selecting a KIS sample if the design parameters are fully documented. The task for the KIS survey will be to design a subsampling procedure that will produce a sample in line with KIS requirements. This will not always be possible. However, the larger the master sample is in relation to the desired KIS subsample, the more flexibility there will be for developing a subsampling design.

A key question with a preexisting sample is whether the listing of dwellings/households is still current or whether it needs to be updated. If updating is required, use of a preexisting sample may not be economical.

The potential *advantages* of using a preexisting sample are: 1) economy, and 2) increased analytic power through comparative analysis of two or more surveys. The *disadvantages* are: 1) the problem of adapting the sample to KIS requirements, and 2) the problem of repeated interviews with the same household or person in different surveys, resulting in respondent fatigue or contamination. One way to avoid this last problem is to keep just the primary sampling units and reselect the households for the KIS survey.

4.4 Stratification

Stratification is a process by which the survey population is divided into subgroups or *strata* that are as homogeneous as possible using certain criteria. The principal objective of stratification is to reduce sampling error. In a stratified sample, the sampling error depends on the population variance existing *within the strata* but not *between strata*. For this reason, it is important to create strata with low internal variability (or high homogeneity). Another reason for stratification is that, where marked differences exist between subgroups of the population (e.g., urban vs. rural areas), stratification allows flexible sample designs separate for each subgroup.

Strata should not be confused with survey domains. A survey domain is a population subgroup for which separate survey estimates are desired (e.g., urban areas/rural areas). A stratum is a subgroup of homogeneous units (e.g., subdivisions of an administrative region) in which the sample may be designed differently and is selected separately. Survey domains and strata could be the same but they need not be. For example, survey domains could be the first-level stratum in a multi-level stratification. A survey domain could consist of one or several lower-level strata.

Explicit stratification involves deliberate sorting and separating of the units into strata; the sample is then selected independently within each of the specified strata. *Systematic sampling* of units from an ordered list (with a fixed interval between selected units) can also achieve the effect of stratification. This is called *implicit stratification*.

Stratification can be single-level or multi-level. Single-level stratification is used to divide the population into strata according to certain criteria. Multi-level stratification is used to divide the population into first-level strata according to certain criteria and then to subdivide the first-level strata into second-level strata, and so on. A typical two-level stratification is region crossed by urban-rural stratification. A KIS survey is usually multi-level stratified.

If the survey area is large, KIS should use explicit stratification to create separate survey domains for urban and rural residence. Where data are available, explicit stratification could also be done on the basis of socioeconomic zones or more directly relevant characteristics such as the level of female literacy or the presence of health facilities in the areas. These kinds of information could be obtained from administrative sources. Within each explicit stratum, the units can then be ordered according to location, thus providing implicit geographic stratification.

Finally, stratification should be introduced only at the first stage of sampling. At the dwelling/household selection stage, systematic sampling is used for convenience; however, no attempt should be made to reorder the dwelling/household list before selection in the hope of increasing the implicit stratification effect. Such efforts generally have a negligible effect.

4.5 Sample Size Determination

The issue of sample size determination is only partly a technical one. Under the same survey conditions, *the larger the sample size, the better the survey precision and the more elaborate the analyses that can be sustained*. The challenge in deciding on the sample size for a survey is to balance the demands of analysis with the capability of the implementing organization and the constraints of funding.

An appropriate sample size for a KIS is the minimum number of persons (e.g., women age 15-49 or children under age five) which will allow core indicators to be calculated with a reasonable level of precision. Attaining a reasonable level of precision is usually not a problem for most indicators at the level of entire survey area. Thus, apart from survey cost considerations, *the total sample size depends on the desired precision at the domain level and the number of domains*.

If the same *relative margin of error (RME)*² is desired for all domains, the domain sample size depends on the variability and the size of the domain. The basic formula used in the calculation of the sample size is given by:

$$n = \frac{Deft^2 (p^{-1} - 1)}{\alpha^2}$$

where n is the number of individuals; p is the estimated prevalence rate or proportion; α is the relative margin of error expected; $Deft$ is the design effect³.

² The *RME* of an estimator is the ratio of its absolute error over its estimated value. This measure is independent of the scale of the parameter to be estimated and therefore a unique *RME* can be used for all indicators. The relationship between the half-length of the confidence interval (with a confidence level of 95 percent) and the *RME* is: $P^* RME$ is the half-length of confidence interval for P . For example, for $RME=0.15$ and $P=0.384$, the half-length of the confidence interval is 0.058.

Clicking on the sample size determination icon in Figure 4.1 at the end of the chapter will bring up a spreadsheet that will facilitate the calculation of the sample sizes required at varying levels of precision. Table 4.1 illustrates the use of the sample size determination spreadsheet in calculating the domain sample sizes required to estimate indicator “*the proportion of currently married women age 15-49 who are currently using any contraceptive method*” at different levels of RME. For the example in Table 4.1, the following parameters⁴ were entered in the spreadsheet: (1) the assumed contraceptive rate in the domain, (2) the design effect (Deft)⁵; (3) the number of target individuals (currently married women 15-49) per household; and the individual and the household response rates.⁶ Table 4.1 shows, 1,783 households would have to be selected to provide an estimate of the CPR at an RME of 15 percent in this particular domain.

The total sample size for a survey with several domains would be equal to the sum of the sample sizes obtained in the above table for each domain. If the indicator level and the precision are the same for all domains, then the total sample size is the sample size calculated for one domain multiplied by the number of domains. Using the example in Table 4.1, if the desired RME is 15 percent, the total sample size for a survey having five domains with approximately the same CPR level of contraception prevalence would be 8,951 households, i.e., a sample of 1,783 households would be required in each domain.

Table 4.2 presents a similar example for the indicator “*the proportion of children under five who received an ORT for diarrhoea in the two weeks preceding the survey*”. In this case, the estimated number of target individuals per household (children under five who had diarrhoea in the last two weeks preceding the survey) is much smaller than the number of currently married women per household in Table 4.1. As a result, to achieve an RME of 15 percent, we would need to select 7,143 households, which is considerably larger than the sample size that we saw in Table 4.1 would be required to estimate the CPR for women at the same level of precision.

This comparison illustrates that for a multi-indicator survey, the sample size required to provide estimates at a given level of precision may vary considerably. Thus, the choice of a main indicator that will be used for the sample size determination is crucial. The main indicator should have demographic importance, at least a moderate prevalence value, and reasonable population coverage, i.e., the average number of target individuals per household should not be very low.

³ For cluster surveys, a 2-step process is commonly used to determine sample size. First, an initial sample size is determined by ignoring the clustering effect. Second, a final sample size is calculated by multiplying the initial sample size by the quantity (Deft)².

⁴ These parameters must be estimated, either based on information from prior surveys or administrative records.

⁵ In the sample size determination spreadsheet, a default value of 1.5 is set for *Deft* if not specified.

⁶ The final sample size takes non-response and finite population correction into account. The number of households needed is converted from the number of individuals based on the number of target individuals per household and the household non-response rate. If no response rate is specified, the template calculates the net sample size.

Table 4.1 Sample size requirements for estimating the contraceptive prevalence rate at varying RME levels

Total target population (individual)					
Estimated prevalence rate (proportion)				0.384	
Estimated design effect (Deft) p				1.710	
Number of target individuals per household				0.574	
Individual response rate				0.940	
Household gross response rate				0.868	
Relative Margin of Error	Sample size household	Sample size individual	Expected STDE	95% confidence limits	
				Low-limit	Up-limit
0.20	1002	499	0.038	0.307	0.461
0.19	1112	554	0.036	0.311	0.457
0.18	1237	616	0.035	0.315	0.453
0.17	1387	691	0.033	0.319	0.449
0.16	1566	780	0.031	0.323	0.445
0.15	1783	888	0.029	0.326	0.442
0.14	2046	1019	0.027	0.330	0.438
0.13	2371	1181	0.025	0.334	0.434
0.12	2784	1387	0.023	0.338	0.430
0.11	3312	1650	0.021	0.342	0.426
0.10	4007	1996	0.019	0.346	0.422
0.067	8926	4447	0.013	0.358	0.410

Note: If not entered above, the default value of Deft is set to be 1.5..

Table 4.2 Sample size requirements for estimating the prevalence of ORT among children under age five who had diarrhea in the two weeks preceding the survey

Total target population (individual)					
Estimated prevalence rate (proportion)				0.292	
Estimated design effect (Deft) p				1.219	
Number of target individuals per household				0.110	
Individual response rate				0.940	
Household gross response rate				0.868	
Relative Margin of Error	Sample size household	Sample size individual	Expected STDE	95% confidence limits	
				Low-limit	Up-limit
0.20	4012	383	0.029	0.234	0.350
0.19	4452	425	0.028	0.237	0.347
0.18	4965	474	0.026	0.239	0.345
0.17	5562	531	0.025	0.242	0.342
0.16	6274	599	0.023	0.245	0.339
0.15	7143	682	0.022	0.248	0.336
0.14	8191	782	0.020	0.251	0.333
0.13	9510	908	0.019	0.254	0.330
0.12	11155	1065	0.018	0.257	0.327
0.11	13281	1268	0.016	0.260	0.324
0.10	16056	1533	0.015	0.263	0.321
0.300	1791	171	0.044	0.204	0.380

Note: If not entered above, the default value of Deft is set to be 1.5..

In practice, decisions about domain sample sizes are often dictated by budget constraints. The total sample size for a survey is often fixed according to available funding, and the sample is allocated to each domain. The following section describes procedures that may be used to allocate the sample at the domain level once the total sample size has been determined. Regardless of the method used for allocation, the calculation of domain sample size is a useful exercise since it informs us about the precision we may achieve in each domain with a given sample size.

4.6 Sample Allocation

Once the total sample size has been fixed, we need to appropriately allocate the sample to the various domains or, within domains, to the strata of interest. Because the KIS is a multi-purpose survey, a proportional allocation of the sample is recommended if the strata are not too different in size. If the domain or strata sizes are very different, a *power allocation* with an appropriate power value may be used to guarantee sufficient sample size in small strata. A power value of 1 gives proportional allocation, while a power value of 0 gives *equal size allocation*, and a power value between 0 and 1 gives an allocation between proportional and equal allocation.

Clicking on the sample allocation icon in Figure 4.1 at the end of this chapter will bring up a spreadsheet that will facilitate the process of allocating the sample. Table 3 illustrates the use of the spreadsheet in proportionally allocating a sample of 9,000 individuals across 5 domains. The proportions shown in the table represent the total target population in the domain (stratum) over the total target population.

Table 4.3 Sample size allocation—proportional allocation

Sample Size	9000	
Power Value		
Domain/Stratum	Proportion	Allocation
1	0.161	1449
2	0.301	2709
3	0.222	1998
4	0.048	432
5	0.268	2412
Total	1.000	9000

The example in Table 4.3 illustrates the considerable variability in the sample sizes that can result using proportional allocation; the expected sample size varies from 432 in Domain 4 to 2,709 in Domain 2. Table 4.4 provides an example in which the power value has been adjusted to ensure a minimum sample size of at least 1,000 in each domain. In this case, the small domains are oversampled compared with a proportional allocation. Oversampling some small domains is frequently necessary if domain-level tabulation and precision are required.

Table 4.4 Sample size allocation—power allocation

Sample Size	9000	
Power Value	0.400	
Domain/Stratum	Proportion	Allocation
1	0.161	1710
2	0.301	2196
3	0.222	1944
4	0.048	1054
5	0.268	2096
Total	1.000	9000

4.7 Sample Take

After the total sample size has been fixed and before the selection of the EAs, we must decide on the number of households to be selected in each EA and then calculate the total number of EAs that need to be selected. The optimum number of households to be selected per EA depends on the variables under consideration, the size of the EA, and the relative sampling cost per EA and per household.

A larger sample size within each EA can reduce survey field costs, but it can also reduce the survey precision if the households are very similar with respect to the variable(s) under consideration in the survey. Because EAs usually consist of geographically coherent households, experience shows that there is often considerable homogeneity among the households within an EA (Aliaga and Ren, 2004). To reduce the potential impact of this homogeneity, it is recommended that a large sample take within each EA should be avoided. For a moderately average EA size of 100-300 households, the optimum sample size ranges from 20 to 40 households (Aliaga and Ren, 2004). For details of size of sample taken per EA, refer to the *DHS Sampling Manual* (Macro International, 1996) and Aliaga and Ren, 2004.

4.8. Sample Selection: a Two-Stage Procedure

For the KIS, a two-stage systematic sampling procedure is recommended. In the first stage, every EA in the survey area is assigned a measure of size equal to the total number of households (population) in the EA. In each domain (stratum), a sample of EAs with a predetermined sample size is then selected independently with probability proportional to this measure of size. In the selected EAs, a listing procedure is performed such that all dwellings and households are listed. This procedure is important for *correcting errors existing in the sampling frame*, and it *provides a sampling frame for household selection* (see details below and in Macro International, Inc. 2004). After a complete household listing is

conducted in the EAs, a fixed number of households are selected by equal probability systematic sampling in the selected EAs.

Clicking on the PPS icon in Figure 4.1 at the end of this chapter will bring up a spreadsheet that will facilitate the process of selecting sample clusters (EAs). Table 4.5 below gives an example of the information input table the PPS selection procedure.

Table 4.5 Information input table for PPS selection

Random (0, 1)*						Col name of strata*		
Stratum num								
Stratum size								
St Sample size								
Stratum num								
Stratum size								
St Sample size								
Stratum num								
Stratum size								
St Sample size								
Stratum num								
Stratum size								
St Sample size								
Col name of Dom/Reg			Col name of urbrural			Col name of PSU size		
Total number of strata			Total sample size			# of Diff PSU selected		
Dom/Reg name/code	Urban/rural	PSU Size	Stratum number	Select Proba	# of times Select	Stratum size	Stratum sam-size	Meaure size-strat

4.9 Segmentation, Mapping and Listing

After the EAs are selected, a complete listing of dwellings/households in the EAs is necessary before the selection of households. Before the listing can be carried out, it may be necessary to further segment EAs with very large populations. If possible, it is recommended that segments of approximately equal size be created. Typically, about 200 households is an appropriate segment size if 25-30 households are to be selected in the entire EA.

Segmentation becomes progressively more difficult as segments become smaller because there are not enough natural boundaries to delineate very small segments. Moreover, concentration of the sample into smaller segments increases the sampling error. Because neighbours' characteristics are correlated, a smaller segment captures less of the variety existing in the population, which leads to less efficient sampling. There is a point beyond which it is not useful to attempt further segmentation. As a general rule, the average segment size should not be less than 100 households.

In some cases, the census maps may be accurate enough for the work of segmentation to be done in the office. More typically, a field operation may be needed to map and segment oversized EAs. If size measures (e.g., the number of households) are required, these can be obtained at the same time using a quick count. To better control the fieldwork, it is recommended that only the fieldwork coordinator or team supervisor has the authority to decide which EA should be segmented and how many segments will be created in the EA.

Selection of the sample segment in each segmented EA is the next step. It is important to prevent biased selection so it is recommended that materials be returned from the field and the selection be done in the survey office. If the selection is done in the field, clear instructions on how to select the segment should be given to the team doing the segmentation in the field, together with necessary parameters (i.e., the random number). A probability proportional to segment size selection is recommended. Furthermore, control procedures should be introduced to guard against bias. For more details of the segmentation operation, see Macro International, Inc. 2004.

The next step is mapping and listing. The mapping and listing operation consists of visiting each of the selected clusters, recording on listing forms a description of every structure together with the names of the heads of the households found in the structure, and drawing a location map of the cluster as well as a sketch map of the structures in the cluster. The listing operation represents an appreciable field cost, but there is no reliable method by which it can be avoided. The listing operation represents one of the most important bias correction procedures in the survey, especially when the sampling frame is out-of-date.

Experience shows that more, rather than less, attention to the quality of listing operations is required if serious biases are to be avoided. In particular, the combination of listing, sampling, and interviewing into a single operation, conducted by the interviewer while moving over the sample area, is unworkable. Even less acceptable is the attempt to avoid listing altogether by having interviewers create clusters as they go along, or select a sample at fixed intervals during a random walk up to a predetermined quota. Such methods are designed to eliminate conscious choice in selection, but they fail to meet the requirement that the sample be selected in such a way as to give a known and nonzero probability to every potential respondent. Essentially, these methods represent a false economy. *It is more efficient to reduce the sample size and retain the listing operation.*

4.10 Household Selection

Once the mapping and household listing operation is completed, the household lists should be sent to the central survey office for the selection of households. The recommended household selection procedure is equal probability systematic sampling. This procedure consists of selecting the sample households from the listing with a random start by the following criteria:

Let L be the total number of households listed in the cluster; let $Random$ be a random number between $(0, 1)$; let n be the number of households to be selected in the cluster; let $I = L/n$ be the sampling interval.

- (1) The first selected sample household is k (k is the serial number of the household in the listing) if and only if:

$$(k-1)/L < Random \leq k/L$$

- (2) The subsequent selected households are those having serial numbers:

$$k + (j-1)*I, \text{ (rounded to integers)}$$

for $j = 2, 3, \dots, n$;

It is important to note that the *Random* numbers should be different and independent from cluster to cluster.

Clicking on the 'HHs Selection' icon in 4.1 at the end of the chapter will bring up a spreadsheet that will facilitate the household selection. Table 4.6 illustrates a part of the template. When household listing results are entered, the selected households will appear automatically in the designated places.

Table 4.6 Template for household selection

Cluster ID						Cluster selection proba	Nbr HHs Listed	Nbr HHs Selected	Select interval	HHs design weight	Random (0-1)	1	2	3	4	5
x	x	x	x	x	x	0.01435	125	25	5.00	348.43	0.60281	4	9	14	19	24
x	x	x	x	x	x	0.02147	160	25	6.40	298.09	0.95636	7	13	19	26	32
x	x	x	x	x	x	0.01945	134	25	5.36	275.58	0.57949	4	9	14	20	25
x	x	x	x	x	x	0.02044	90	25	3.60	176.17	0.40303	2	6	9	13	16
											0.93286					
											0.84306					

Though an equal probability systematic sample is easy to select, centralization of the household selection is necessary so that the completeness of the household listing operation can be assessed by experienced survey staff. Discrepancies between the expected and the listed number of households must be evaluated. Problem areas should be revisited. Sampling fractions could also be readjusted so as to give the expected number of households. In cases where it is not feasible to centralize household selection, especially when regional household listing teams are employed and travel is difficult, supervisors can be trained to do the selection in the field. However, in this situation, the evaluation of the quality may not be possible.

After the selection of households, interviewing teams will be sent to the clusters and interviewers will be assigned selected households to interview. The interviewer must visit only the households he/she has been assigned, and does not have the right to change/replace a previously selected household. Any unusual circumstances (dwellings not found, destroyed, or vacant) must be properly documented and reported.

4.11 Sampling Weights

The KIS sample will not generally be self-weighting because a self-weighting design is very complicated and depends on projections of the target population in small areas. Since the KIS sample is a two-stage stratified cluster sample, sampling probabilities will be calculated separately for each sampling stage and for each cluster. We use the following notations:

P_{1hi} : first stage's sampling probability of the i^{th} cluster in stratum h

- P_{2hi} : second-stage's sampling probability within the i^{th} cluster (households)
 P_{hi} : overall sampling probability of any households of the i^{th} cluster in stratum h

Let a_h be the number of clusters selected in stratum h , M_{hi} the number of households according to the sampling frame in the i^{th} cluster, and $\sum M_{hi}$ the total number of households in the stratum h . The probability of selecting the i^{th} cluster in stratum h is calculated as follows:

$$P_{1hi} = \frac{a_h M_{hi}}{\sum M_{hi}}$$

Let b_{hi} be the proportion of households in the selected segment compared to the total number of households in EA i in stratum h if the EA is segmented, otherwise $b_{hi} = 1$. Let L_{hi} be the number of households listed in the household listing operation in cluster i in stratum h , let g_{hi} be the number of households selected in the cluster. The second stage's selection probability for each household in the cluster is calculated as follows:

$$P_{2hi} = \frac{g_{hi}}{L_{hi}} \times b_{hi}$$

The overall selection probability of each household in cluster i of stratum h is the product of the selection probabilities:

$$P_{hi} = P_{1hi} \times P_{2hi}$$

The sampling weight for each household in cluster i of stratum h is the inverse of its selection probability:

$$W_{hi} = 1 / P_{hi}$$

This weight needs to be adjusted for household non-response. The adjusted weight will be further normalized for the whole sample so that the total number of weighted cases is equal to the number of unweighted cases. This normalized household weight is the gross sample weight for individuals living in the households in the same cluster. This weight is further adjusted for individual non-response and then normalized to get the final individual sample weight. It needs to be pointed out that the normalized weights are valid for estimation of proportions and means at any aggregation levels, but not valid for estimation of totals.

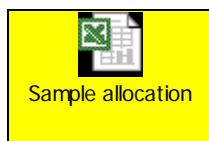
Clicking on the standard weight icon in Figure 4.1 at the end of this chapter will bring up a spreadsheet that will facilitate the process of calculating the weights.

Figure 4.1 Excel templates for sampling

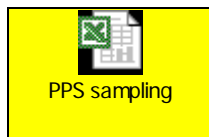
1. Sample size determination



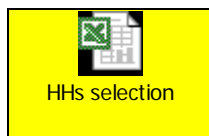
2. Sample allocation



3. Stratified systematic sampling with probability proportional to size (PPS)



4. Household selection procedure



5. Standard weight calculation