Integrated Framework for Household Survey

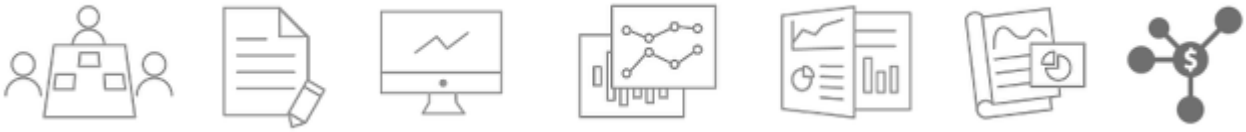
IFHS: A toolkit to facilitate design, collection & analysis.

Last update : 3 March 2019

Why an "Integrated Framework"?

The kit is a comprehensive set of tools to guide survey teams through every step of a multi-sectoral needs assessment done through a **household survey** – from overall planning, design and data collection in the field to data processing, analysis, interpretation, documentation and dissemination. Because it addresses the specific challenges related to household surveys, it fits more for protracted situations and it **complements** other approaches such as the UNHCR Tool for Participatory Assessment in Operations, the Needs Assessment for Refugee Emergencies (NARE) Checklist for Refugees operations and the Rapid Protection Assessment Tools (RPAT) for Internally Displaced Persons (IDPs) operations. The kit is also organised to ensure that all findings and recommendations will allow to inform prioritization between all potential interventions and subsequent resources allocation in terms of programme, through the following elements :

- Produce **impact indicators** that are used to assess the conditions of persons of concern (PoCs) over time and to support programme design;
- Support the **analysis of protection risk and multi-dimensional vulnerabilities** and the discovery of clusters of individuals with similar profile;
- Allow for the development of **targeting models** for both response (e.g. cash) and prevention activities (e.g. protection);
- Provide basis for **public advocacy** on issues faced by the population group.



This toolkit supports the technical implementation of already existing guidelines and guidance documents, namely the Guidelines for Integrating Gender-Based Violence Interventions in Humanitarian Action, the Child Protection Rapid Assessment Toolkit, the Joint IDP Profiling Service Essential Toolkit-JET, the Heightened Risk Identification Tool, The Operational Guidance and Toolkit for Multipurpose Cash Grants, etc. The proposed technical implementation approach was inspired by the UNHCR Standardised Expanded Nutrition Survey and the UNICEF Multiple Indicator Cluster Surveys.

Content

On the top of the summary narrative guidelines below, the toolkit is organised around four key technical components that facilitate the easy replication and customisation of an assessment from one operation to another. Re-using tested platforms, questions and report formats, technical staff in charge will save significant amount of time.

These elements are:

NOTE

- A pre-organised Master list of baseline Indicators Library in order to leverage good practices and enforce core questions during design of the assessment form.
- A KoboToolBox secure server to be used for data collection. This server uses the xlsform developed during the design phase.
- Multiple R statistical analysis scripts to clean & analyse data, and then to generate automatically standard report and to facilitate the creation of presentation slides and infographics.
- A Github Repositories to facilitate collaborative analysis between operations and generate a knowledge base. Github is also used for the maintenance of the toolkit itself. Suggestions for this toolkit can be posted here.

For whom?

The targeted audience of the toolkit is *technical staff tasked to work on multi-sectoral needs assessments*: assessment focal points, information management officers and data scientists. The toolkit is intended for protracted situations, **after** emergency assessments and secondary data review have been conducted. With a bit of configuration at the beginning of the process, the toolkit will ensure that minimum data quality standards are enforced and I allow for quick generation of results. . In addition, because of the standard data format presented through the toolkit, it will be possible to obtain comparable results from different operations.

Disclaimer

The toolkit is a *collaborative effort*: if you have suggestions, please share them through this link.

The Toolkit is a *work in progress*: if you identify issues, please share them through this link.

Table of Contents

Assessment methodology & form design

Planning

Methodology

Form

Data collection

Preparing for fieldwork

Using KoboToolBox

Fieldwork manual

Analysis & dissemination

Analytics Steps

Analysis Process

Communication

Credits

Large parts of the toolkit are extracted from other existing guidelines referenced above.

Assessment Project Document

NOTE

A multi-sectoral needs assessment must be carefully planned to maximize efficiency and ensure actionable results. The first – and perhaps most important – step of this plan is to ensure all elements of the process are documented. Keeping this in mind from the beginning will prevent a frantic dash later on, when reporting to donors or answering evaluations.

The project document should: * clearly state the assessment objectives; * outline the governance structure, including roles and responsibilities; * define the methodology to be employed (population sample, tools, etc.) * provide an overview of staffing needs, budget, and timeline.

The document can be updated throughout the exercise.

Background and Objective

In UNHCR's work around the world, multi-sectoral needs assessments are designed to collect statistically sound, internationally comparable estimates of key indicators and analysis that are used to assess overall situation of PoCs, and to shape and prioritize appropriate interventions. Data should always be collected for a clear purpose, and only when necessary.

Governance Structure

From the outset, it's important to define roles and responsibilities. This is true of single-agency and joint assessments. In recent years, joint assessments have become more of a reality, and they are an important element of the Grand Bargain commitment to coordinated needs assessments. Despite a number of benefits, joint assessments can at times be challenging. It is important to establish a strong governance structure from the beginning, often outlined in a Terms of Reference (ToR). Even single-agency assessments are often conducted in collaboration with other parties, such as the government, external contractors, etc. A ToR document for any assessment should:

- Give the name and type (government agency or other agency) of implementing agency.
- Provide overview of Memorandum of Understanding (MoU) (Parties, critical components affecting survey planning, etc.)
- Give the names and affiliations of those who will be responsible for the management, technical work, and coordination activities. Include the survey coordinator, the sampling expert, and data processing expert assigned from the implementing agency, as well as others, if applicable. If applicable, regional experts/consultants together with their respective responsibilities should also be included.
- Describe the roles and contributions of national and international stakeholders and funding agencies.
- Describe the status, composition and roles and responsibilities of the Steering and Technical committees.
- Provide other details on the governance structure and human resources as needed.

Methodology

Data analysis plan

A data analysis plan should be devised in the early stages of assessment planning. Rather than deciding on questions first, indicators must be determined in line with the information needs. Starting with the data analysis plan will avoid unpleasant surprises at the end of data collection, when you realise a certain

important question might have been left out! Additionally, if the data analysis plan is linked to the tool from the outset, this will save time in data processing and analysis.

Tool

While UNHCR does not have a standardized tool for multi-sectoral needs assessments, many questionnaires used in different countries are similar to each other, primarily due to cross-pollination of staff and regional experts. Tools can be easily adapted according to operational needs. Once your indicators are defined, questions can be chosen, preferably from the IHSN.

UNHCR's corporate tool for data collection is Kobo. The assessment focal point and other relevant staff should create an account on kobo.unhcr.org

Provide information on the plans for the translation and back-translation of the questionnaires into local languages and plans for pre-testing the questionnaires. Indicate that the pre-test results will be compiled in a report, and that the results of the pre-test will be used to further modify, customize, and finalize the questionnaires.

Sample Design

Sample design is crucial for usability of results. A badly designed or implemented sample can result in the findings not being applicable in the way that was originally intended, be it at geographical, demographic or any other level. Support for sample design can be requested from regional offices, and HQ (more specifically FICCS). Do not hesitate to reach out for support!

UNHCR has also created a Sampling Decision Assistant to help in getting a general idea of how many households should be surveyed in your assessment:

Any sampling strategy, however, should still be validated by technical experts prior to implementation.

Documenting the sampling strategy is also very important for posterior use of the dataset in other studies. In this section, under separate sub-headings, as appropriate, describe: * The Type of sampling design (Rationale for sampling design explained) * Definition of unit (case/household) used in the assessment * Sample size, including the expected numbers of households, women, men and other demographic characteristics as appropriate. * How the sample size was calculated, including the indicators used for the calculation of the sample size * The level of disaggregation sought for reporting * What sample frame will be used and if the sample frame needs to be updated, plans for mapping, listing and household selection

Staffing needs

UNHCR commonly relies on partners/third parties for large data collection exercises. This section of the

document can be completed jointly, with each party submitting inputs as appropriate.

In this section, under separate sub-headings as appropriate, describe:

- Plans for recruitment of fieldwork staff, including details of the type of personnel (interviewers, data entry, supervisors, measurers, data entry clerks), their education/background, sex, numbers etc.
- Timing of training
- Length of training
- Methodology and content of training
- Profiles of trainers
- How training will be organized – central location, in separate districts, including how standardization will be ensured if not central location

Fieldwork

In this section, under separate sub-headings as appropriate, describe:

- Timing of fieldwork, constraints on timing of fieldwork
- Team composition, including numbers
- Expected duration of fieldwork and how the duration was calculated
- Plans for monitoring data collection and fieldwork supervision as well as plans for handling questionnaires for data entry
- Fieldwork logistics

Budget

In this section, under separate sub-headings as appropriate, describe:

- Expected total cost of the survey
- Breakdown of total cost by budget line items
- Amount of funding secured and funding source(s)
- Amount of extra funding needed, including plans, if any, on how the funding shortfall will be secured

Timeline for Tasks

- Identify survey coordinator, survey personnel, and plan survey; establish steering and technical committees
- Adapt and pre-test questionnaires; translate questionnaires and manuals

- Carry out sampling and household listing; order scales, boards, salt test kits, and GPS equipment
- Complete logistical arrangements
- Select and train fieldwork personnel (interviewers, editors, measurers, and supervisors)
- Conduct pilot study and collect data
- Complete data processing, including secondary editing
- Prepare summary findings report and final report, and disseminate widely; prepare survey archive

Memorandum of Understanding

IMPORTANT

Whether it is for the data collection or the analysis, protection assessment are often done in partnership. To avoid confusion and misunderstanding within the process, it is recommended to establish a clear Memorandum of Understanding.

Implementation Model

According to the particular context, the assessment might be implemented using one of the following models:

- Assigning full data collection responsibility to different agencies in different geographic areas.
- Pooling human and logistical resources centrally.
- Delegating data collection responsibilities to one or more NGO partners, ideally local NGOs, through a project agreement.

Regardless of the model chosen, it is important to observe four principles:

- The objectives and methodology of the assessment need to be decided by consensus among participants. Agencies who assume particular data collection responsibilities do not acquire a privileged say in choosing indicators, sites or data collection methods.
- Participants contribute resources to the implementation of the asessment, to the measure of their capacities and possibilities.
- Common standards and understanding of assessment questions need to be agreed and maintained.

- Clear focal points need to be appointed for each participating organisations with operational responsibility for data collection.

Field Level Memorandum of Understanding

The following is an outdated example that might provide some inspiration for the initial drafting of the memorandum. Any memorandum needs to be cleared by the Bureau before signature.

Scope of Works

The following is an example of MoU.

> - *The following Memorandum of Understanding between the United Nations High Commissioner for Refugees and* `Partner Name` *is not attempt to repeat the basic principles of already existing global MOU, including descriptions of the agency responsibilities towards various populations, but will instead highlight the specific areas where close cooperation will be taking place.*
> - *This MoU covers the geographic Coverage for* `population group`*. All areas will be covered based on* `sampling methodology`
> - *UNHCR and* `Partner Name` *will agree on a joint questionnaire subsequently used for all data collection related to the profiling exercise, through mobile data collection devices and a* `methodlogy` *(for instance: combination data collection methodologies: direct observation, key informant interviews and as means of triangulation in areas with a high concentration of* `population group`*, carry out Focus Group Discussions).*
> - *All data collectors will be selected jointly and need to sign the code of conduct. They will receive a ½ day training on the principles and objectives of the code of conduct. In addition, data collectors will be trained at a minimum in: interviewing techniques, protection principles, data protection (confidentiality, informed consent etc.), referral mechanism, usage of mobile devices, the questionnaire, and basic security principles. This introductory training will be conducted before any data collection takes place and will last 3 to 4 days. One day of field testing with the data collectors is foreseen at the beginning of the roll-out. Technical documentation of the profiling exercise will include: data collectors manual, team leader manual, data analysis plan, data*

entry manual and the respective Standard Operating Procedures for the project.

- *The role of the Government from the beginning of this exercise is crucial to ensure a responsible handover at the end of the project and overall ownership by the Government. To this end, the Government will be included in the project through training and in particular Line Ministries in the localities to contribute to data collection.*
- *The questionnaire will be presented to and feedback sought from partners (UN, NGOs and Government) to achieve buy-in and ensure that the questionnaire meets the information needs of partners providing protection and assistance to* `population group`*; however, given the need for timely information delivery, contributions from partners are expected to be provided within a week from the day they receive the draft questionnaire.*
- *UNHCR and* `Partner Name` *will have joint data ownership. The data will be stored on a UNHCR server, while UNHCR and* `Partner Name` *will have both administrative rights to access the data base. UNHCR and* `Partner Name` *will disseminate and/or publish the data collected jointly. None of them will be producing the data alone to publish products /reports /websites etc. under its own logo or other branding.. The design and layout of all information products of this project will be agreed jointly. Data collection will be an ongoing exercise to ensure that information is updated, relevant and timely. At the end of the project the data will be handed over to the Government.*
- *UNHCR and* `Partner Name` *will establish a joint budget and enter into a cost-sharing arrangement to finance this project.*
- *In line with UNHCR's mandate for refugees, UNHCR will be collecting data on refugees residing outside camps which is not part of this MoU.*
- *Nothing in this MOU shall affect the relations of either signatory to its Governing Body, nor the contractual relationship and administrative supervision of UNHCR and* `Partner Name` *to their operational partners.*
- *The implementation of the MOU will be in compliance with the respective administrative and financial rules and procedures of UNHCR and* `Partner Name` *and be subject to the availability of funds.*
- *This MOU will enter into force upon signature and shall be of indefinite duration.*
- *This MOU may be terminated by either party upon 90 days written notice.*

- *This MOU may be modified at any time by mutual consent of the parties.*
- *The Representatives of both organizations will meet when necessary to discuss policy issues and will nominate officers to meet regularly to review strategic and implementation issues of particular interest to both organizations and to propose possible courses of action to address them.*

Terms of Reference for Assessment Focal Point

Background

Insert

- background on current protection situation,
- details on previous assessment or what is known of the protection situation.

Explain why the current assessment is planned and if there is a specific trigger that would indicate a situation change.

Insert any other relevant detail.

Methodology

Survey methodology should be reviewed by UNHCR regional or HQ Infornmatiuon Management Officer **prior to data collection**.

Objectives

The Assessment Focal Point will oversee the multi-sectoral needs assessment for:

- Operation
- Population Group
- Geographic Coverage
- Timeframe

Deliverables

- A final assessment report including recommendations on actions to address the situation is to be submitted at the end of the mission. Results of standardisation tests, details of data cleaning and plausibility checks should be presented in the final report.
- Standardised tables as presented in the multi-sectoral needs assessment Toolkit.
- The findings and major recommendations are to be presented to partners at the mission level (oral presentation and slides).
- The final analysis script.

Reporting

The consultant will report on regular basis to the UNHCR [insert title of UNHCR person responsible], who will have the overall responsibility of the survey.

Time frame

The consultancy will last approximately [insert number of months], starting from [Insert start date].

Qualification & Experience required

The successful candidate will:

- Have a university degree or the equivalent in social science, with a specific competency in humanitarian emergencies.
- Have significant experience in undertaking surveys (design and methodologies, staff recruitment and training, field supervision and data analysis/write up).
- Be familiar with the survey methodology and R statistical language.
- Be fluent in English with excellent writing and presentation skills [insert any other language

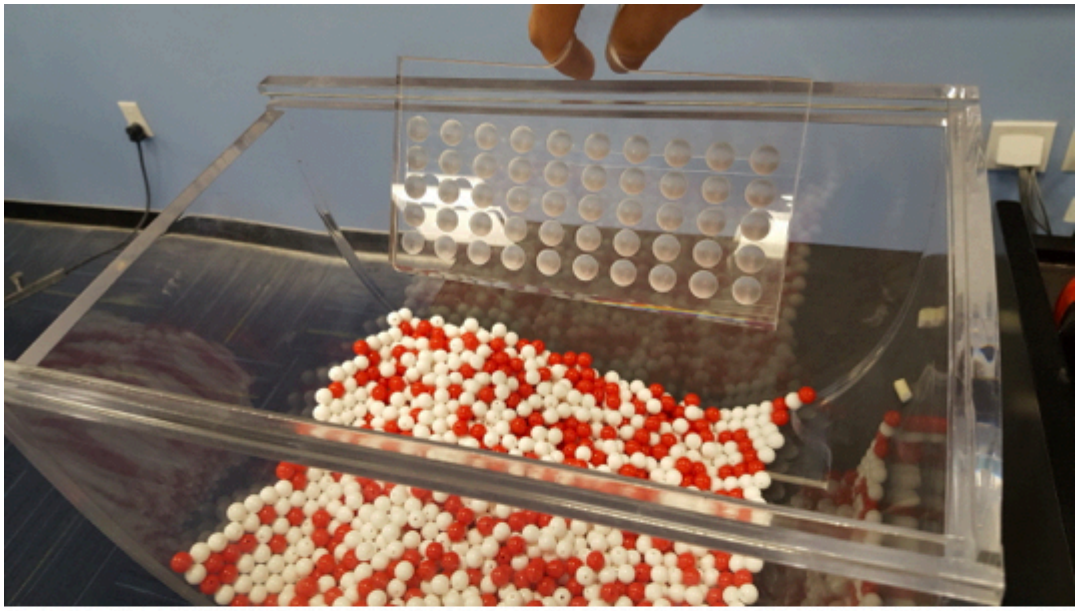
requirements].

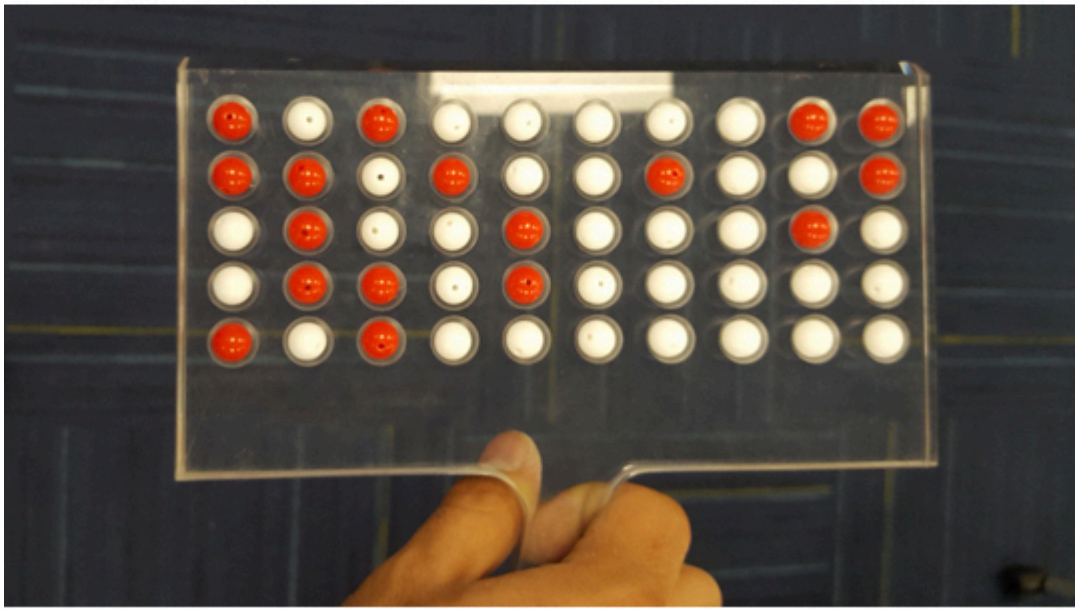
Sampling

IMPORTANT

Sampling strategies are constrained by available budget, field accessibility and time.

hus, the chosen approach for a defined context often reflects a trade-off between representativity of the results, rapid delivery and cost effectiveness.





Sampling strategy

Sampling strategy can be either probabilistic or non-probabilistic. A good introduction can be found here

Non-probabilistic approaches

Non-probabilistic approaches are usually **favored during the emergency phase** where both time and field

access represent the main challenge.

Convenience sampling

A frequently used method in emergency situations, it relies on sampling those respondents who are easiest to access.

Practically speaking those couldd be either: * Key Informants willing to be interviewed.

- Individuals or household among those who have settled along roadsides, or who present themselves to administrative center of the returnee settlement or the assistance desk, etc.
- **Advantages**: Easy and quick to implement, especially when time and access are the main constraints.
- **Disadvantage**: The danger with this type of data collection approach is that it will often lead to biased results as the sample may not be representative of the majority, i.e. those with the most resources or power are often the ones who settle in the most easily accessible areas.

Snowball sampling

Snowball sampling (or chain sampling, chain-referral sampling, referral sampling) is a non-probability sampling technique where existing study subjects recruit future subjects from among their acquaintances. This technique is subject to numerous biases. For example, people who have many friends are more likely to be recruited into the sample.

- **Advantages**: Useful when targeting specific groups that might be difficult to reach (hidden population).
- **Disadvantage**: This approach might underweight the most vulnerable individuals.

Purposive sampling

It is based on previous knowledge about who might be able to provide valuable or specific information. It uses the judgement of community representatives, project staff or assessors to select typical locations and/or informants. The sampling of children or women, for example, is a type of purposive sampling.

Purposive sampling can also be done through Key Informant.

- **Advantages**: Moderately rigorous if well and clear criteria for sampling are followed. Useful when targeting specific groups of affected population or specific affected areas. Less time consuming and less expensive than representative sampling.
- **Disadvantage**: Generalisations are biased and not recommended. Samples are not representative of population due to subjectivity of respondents.

The risk of losing certain componnent of the population can be addressed by defining strata within the purposive sample.

In the case of Desk interview or key Informant, the more observations the better. Some kind of credibility scoring can be obtained for each locations based on a review of the key informant.

Quota sample

A quota sample might be representative of the population (if quotas actually do work, which is not always the case). But a quota sample will never satisfy the strict randomness requirements that statistics require. Only if we are working with a random sample can we make inferences from the sample to the population. In quota samples, there is not sufficient randomness, as the interviewer selects the interviewees actively. Therefore, quota samples cannot be used to reason about the general population.

Probabilistic approaches

Whenever the situation is becoming more **protracted**, probabilistic approaches should be favored. They will allow to generate more reliable results.

Respondent-driven sampling -RDS

A declination of snowball sampling is the Respondent-driven sampling -RDS approach. It combines "snowball sampling" with a mathematical model that weights the sample to compensate for the fact that the sample was collected in a non-random way. As such it can be classified as probabilistic approach. The advantage is that seeds selection is specific and does not require sample frame.

While data requirements for RDS analysis are minimal, there are three pieces of information which are essential for analysis (RDS analysis CANNOT BE PERFORMED without these fields for each respondent):

- Personal Network Size (Degree) - Number of people the respondent knows within the target population.
- Respondent's Serial Number - Serial number of the coupon the respondent was recruited with.
- Respondent's Recruiting Serial Numbers - Serial numbers from the coupons the respondent is given to recruit others.

A good introduction to the organisation of RDS is in this presentation.

Time-Location Sampling

The Time-Location Sampling (TLS) approach can be used when the goal is to have a representation of

population in movement. The idea and the assumption is to sample persons at locations and at time at which they may be found.

Time-location sampling is used to sample a population for which a sampling frame cannot be constructed but locations are known at which the population of interest can be found, or for which it is more efficient to sample at these locations. As such the approach is likely appropriate when the survey is taking place at a **border**.

More practical guidelines for TLS are available in a dedicated Resource Guide TLS and some application on Border Monitoring for tourism or illegal migrants.

Random sampling

If you need a purely random sample, the size of the sample is a calculation that takes 3 variables:

- Size of the full population. In refugee Context, Data is coming from proGres while in IDP context, data is coming from a Displacement Tracking System.
- Confidence level: for what proportion of the population you want to get the right estimation (usually either 90%, 95% or 99%)
- Error Margin (or confidence interval): How much error are you willing to tolerate for each questions? i.e. + or – your estimated ratio for each questions on the top of the confidence interval (usually either 5%, 2% or 1%)

There are online calculator for this. Alternatively one can use the excel formula from this example

For 400,000 Syrians	**5% error margin**	**2% error margin**	**1% error margin**
90% Confidence level	272	1694	6692
95% Confidence level	384	2387	9379
99% Confidence level	662	4105	15929

For 150,000 Afghans	**5% error margin**	**2% error margin**	**1% error margin**
90% Confidence level	272	1682	6511
95% Confidence level	383	2363	9026
99% Confidence level	661	4036	14937

Usually the decision on the right confidence level and error margin to be selected is also influenced by cost implication and the final usage of the figures that is looked for.

Stratified sampling

You can refer to this Introduction video or this presentation and this one from the WFP VAM.

A stratified random sample can only be carried out if a complete list of the population is available. In stratified sampling the population is partitioned into groups, called strata, and sampling is performed separately within each stratum.

This can be done for the following reasons:

- Population groups may have different values for the responses of interest.
- If we want to improve our estimation for each group separately.
- To ensure adequate sample size for each group.

In stratified sampling designs, it is assumed that:

- stratum variables are mutually exclusive (non-over lapping), e.g., urban/rural areas, economic categories, geographic regions, race, sex, etc.
- the population (elements) should be homogenous within-stratum, and
- the population (elements) should be heterogenous between the strata.

The major task of stratified sampling design is the appropriate allocation of samples to different strata. The different types of allocation methods includes:

- **Equal allocation**: Divide the number of sample units n equally among the k strata. This implies to use "weighted analysis" (disproportionate selection).
- **Proportional to stratum size**: Make the proportion of each stratum sample is identical to the proportion of the population. A major disadvantage of proportional allocation is that sample size in a stratum may be low and provide unreliable stratum-specific results. In terms of analysis, data will be Self-weighted (equal proportion from each stratum).
- Allocation based on **variance differences among the strata** (called Optimal allocation). Optimal allocation minimizes the overall variance for a specified cost, or equivalently minimizes the overall cost for a specified variance. In situations where the standard deviations of the strata are known it may be advantageous to make a disproportionate allocation. Suppose that, we had stratum A and stratum B, but we know that the individuals assigned to stratum A were more varied with respect to their opinions than those assigned to stratum B. Optimum allocation minimises the standard error of the estimated mean by ensuring that more respondents are assigned to the stratum within which there is greatest variation. Stratum variances are usually defined by previous surveys. This approach also implies to use "weighted analysis" (disproportionate selection).
- Allocation based on the **relative cost of each survey record** (called Neyman Allocation). Neyman

allocation is a special case of optimal allocation where the costs per unit are the same for all strata. In this case, the ideal sample allocation allow to maximize precision, given a Stratified Sample With a fixed Sample Size. The ideal sample allocation plan would provide the most precision for the least cost. This implies to sample more heavily from a stratum when the cost to sample an element from the stratum is low, the population size of the stratum is large or the variability within the stratum is large. This approach also implies to use "weighted analysis" (disproportionate selection).

Typically, when developing the stata definition, in case of optimal or Neyman allocation, i.e. when stratea variance are already known through a previous survey, the following objectives can be looked at:

- Find minimum sample size, given a fixed error
- Find minimum error, given a fixed sample size
- Find minimum error, given a fixed budget
- Find minimum cost to achieve a fixed error

Typical workflow to define sample size in case of stratified sampling:

1. Choose the stratification (e.g.regions, district…)
2. Define the population (N) of each strata
3. Decide on key indicator(s)
4. Estimate mean & variance or prevalence of key indicator
5. Decide on precision and confidence level
6. Calculate the initial total sample size (n) according to the budget/time
7. Use simple random sample per strata to select your representative sample

To estimate sample size, you need to know:

- Estimate of the prevalence or mean & STDev of the key indicator (e.g. 30% return intention). Prevalence is the total number of cases for a variable of interest that is **typically binary** within a population divided by its total population. Mean is the expected value of a variable of interest that is **typically continuous** within a prescribed range for a given population (e.g. expenditure per case)
- Precision desired (for example: ± 5%). Precision is the variability of the estimate.
- Level of confidence (for example: 95%). It represents the probability of the same result if you re-sampled, all other things equal.
- Population (only if below 10,000, otherwise it will not influence the required sample size)
- Expected response rate (for example: 90%)
- Number of eligible individuals per household (if applicable)

Stratified sampling can be performed with R. Tutorial scripts are available here.

Post stratification

One can also use weights, computed through a post-stratification process, to get potentially biased surveys, like online surveys, to better fit the underlying population. The only thing that weights can do, is ensure that your sample composition better mimics the general population's characteristics. Weights will never help you if the process governing non-response is part of the puzzle you want to solve.

In a random sample, we define a population, draw from that population at random and then compute and apply weights to align the sample with the population. This weighting is necessary because some people originally sampled might be e.g. harder to reach than others, thereby biasing the sample. Once the post-stratification weights have been applied, the random sample is representative of the population it was drawn from. Statistics gives us a method to tell just how accurately the findings from the sample can be generalized.

Cluster sampling

Cluster sampling is a technique that allows to reduce the surveying budget when **travel cost are important**. Instead of covering a whole territory, the cluster sampling implies to divide the population into separate groups, called clusters. Then, a simple random sample of clusters is selected from the population.

Cluster sampling are therefore not relevant when techniques such as phone interview are used as there's no marginal surveying cost involved with location of interview.

Given equal sample sizes, cluster sampling usually provides less precision than either simple random sampling or stratified sampling.

Different approaches can be used for cluster sampling

- One-stage sampling. All of the elements within selected clusters are included in the sample.
- Two-stage sampling. A subset of elements within each selected cluster is randomly selected for inclusion in the sample.

Sampling with Replacement and Sampling without Replacement

What is replacement?

When a population element can be selected more than one time, we are sampling with replacement. When a population element can be selected only one time, we are sampling without replacement. When we sample with replacement, the two sample values are independent. Practically, this means that what we get on the first one doesn't affect what we get on the second. Mathematically, this means that the covariance between the two is zero. In sampling without replacement, the two sample values aren't independent. Practically, this means that what we got for the first one affects what we can get for the second one. Mathematically, this means that the covariance between the two isn't zero.

With or without?

In small populations and often in large ones, sampling is typically done “without replacement”, i.e. , one deliberately avoids choosing any member of the population more than once.

Less commonly, sampling can also be conducted with replacement. This allows to address low response rate.

For a small sample from a large population, sampling without replacement is approximately the same as sampling with replacement, since the odds of choosing the same individual twice is low. This can be measure by calculating the covariance: how much two items’ probabilities are linked together. The higher the covariance, the more the results can be influenced. A covariance of zero would mean there’s no difference between sampling with replacement or sampling without.

The specific case of phone surveys

As explained in this paper, bias may be introduced into population estimates through telephone surveys, however, by the exclusion of non-telephone households from these surveys. The bias introduced can be significant since “non-telephone households” may differ from telephone households in ways that are not adequately handled by poststratification. Many households, called “transients”, move in and out of the telephone population during the year, sometimes due to economic reasons or relocation. The transient telephone population may be representative of the non-telephone population in general since its members have recently been in the non-telephone population.

Sample Weight

Over-sampling in regions with small populations ensures that they have a large enough sample to be representative. Under-sampling is done in regions with large populations to save costs. Sample weights are mathematical adjustments applied to the data to correct for over-sampling, under-sampling, and different response rates to the survey in different regions.

How are the oversampled/ undersampled areas corrected in data analysis?

The samples are designed to permit data analysis of regional subsets within the sample population. When the expected number of cases for some of these regions is too small for analysis, it is necessary to oversample those areas. When the expected number of cases for some of these regions is unnecessarily large, those areas may be undersampled to accommodate logistical or budgetary constraints.

During analysis, it is then necessary to “weight down” the oversampled areas and “weight up” the undersampled areas. The developing of the sampling weights has taken this factor into account. Always use the weight variable found in the DHS data set. Even in surveys that come from a self-weighting sample, it is still necessary to use the sampling weights in analysis because the response behavior may

differ by response groups.

What does it mean to normalize the weights?

After the weights are initially calculated, they are normalized, or standardized, by dividing each weight by the average of the initial weights (equal to the sum of the initial weight divided by the sum of the number of cases) so that the sum of the normalized/standardized weights equals the sum of the cases over the entire sample. The standardization is done separately for each weight for the entire sample.

The entire set of household sample weights is multiplied by a constant, thus, the total weighted number of households equals the total unweighted number of households at the national level.

Individual sample weights are normalized separately for women and men. Thus, the total weighted number of women equals the total unweighted number of women, and the total weighted number of men equals the total unweighted number of men. Women and men are normalized separately because all non-HIV calculations are performed on women and men separately. We do not provide survey estimates on the joint population of women and men combined for anything other than HIV prevalence.

Interview

IMPORTANT

Several aspects come into play in the data collection process. The three most crucial aspects include: the cost of the selected data collection method; the accuracy of data collected; and the efficiency of data collection.

In regard to behavioural characteristics, it is generally recognised that face-to-face data deliver the best results, followed by telephone interviews and finally online quota survey. Interactive Voice Response & Short text Message (SMS) Pools are adequate only in very specific cases.

Interview incentives can be very effective to ensure a good response ratio.

Many guidelines are available such as:

- Designing Household Survey Samples: Practical Guidelines
- The Power of Survey Design

Definition of household and relations with UNHR

Registration case

Definition of households and relations with UNHCR Registration case

This concept is explained in details in the (Principles and Recommendations for Population and Housing Censuses.

The concept of household include those persons who live together and have communal arrangements concerning subsistence and other necessities of life, such as eating together. This implies therefore two important arrangements:

- The **household dwelling** concept regards all persons living in a housing unit as belonging to the same household. According to this concept, there is one household per occupied housing unit. Therefore, the number of occupied housing units and the number of households occupying them are equal and the locations of the housing units and households are identical.
- The **housekeeping concept**, that is to say, a person or a group of two or more persons living together who make common provision for food or other essentials for living, with or without combining with any other person to form part of a multi-person household. The persons in the group may pool their resources and have a common budget; they may be related or unrelated persons or a combination of persons both related and unrelated.

Household types

Three types of households can be distinguished:

Nuclear household: defined as a household consisting entirely of a single family nucleus. It may be classified into:

- Married-couple family: With child(ren) or Without child(ren);
- Partner in consensual union (cohabiting partner): With child(ren) or Without child(ren);
- Father with child(ren);
- Mother with child(ren);

Extended household: defined as a household consisting of any one of the following:

- A single family nucleus and other persons related to the nucleus, for example, a father with child(ren) and other relative(s) or a married couple with other relative(s) only;
- Two or more family nuclei related to each other without any other persons, for example, two or more married couples with child(ren) only;
- Two or more family nuclei related to each other plus other persons related to at least one of the

nuclei, for example, two or more married couples with other relative(s) only;
- Two or more persons related to each other, none of whom constitute a family nucleus;

Composite household: like an extended household with the difference of :

- A single family nucleus plus other persons, some of whom are related to the nucleus and some of whom are not, for example, mother with child(ren) and other relatives and non-relatives;
- A single family nucleus plus other persons, none of whom is related to the nucleus, for example, father with child(ren) and non-relatives);
- Two or more family nuclei related to each other plus other persons, some of whom are related to at least one of the nuclei and some of whom are not related to any of the nuclei, for example, two or more couples with other relatives and non-relatives only;
- Two or more family nuclei related to each other plus other persons, none of whom is related to any of the nuclei, for example, two or more married couples one or more of which with child(ren) and non-relatives;
- Two or more family nuclei not related to each other, with or without any other persons;
- Two or more persons related to each other but none of whom constitute a family nucleus, plus other unrelated persons;
- Non-related persons only;

Relations with UNHCR cases as per Registration

The UNHCR case is the equivalent of the nuclear household. UNHCR case number (or ID/Identifiers) are used a basis for large part of the assistance delivery.

When surveying Households, it is important to make connection between the households and the cases:

- Case 1: One single family nucleus which then equals a UNHCR case. In this case both dwelling & housekeeping are de facto shared.
- Case 2: An extended household with two or more than two UNHCR cases. In this case, the surveyor will record if dwelling & housekeeping are effectively shared between cases.
- Case 3: A composite household with two or more than two UNHCR cases, as well as additional members, such as host communities individuals. In this case, the surveyor will record if dwelling & housekeeping are effectively shared between cases, as well as with the members that are not part of the cases.

The main point is to allow for understanding the allocation of expenses (housekeeping & dwelling) between cases that would be grouped together in the same extended or composite household. The allocation could be based for instance on:

- One case covering for all other cases;
- One case covering for non-UNHCR case members;
- Allocation based on number of individuals in each case;
- Allocation based on number of adult individuals in each case;
- Allocation based on number of individuals earning an income in each case, etc.

Comparison of interview approaches

The following is based on Literature review from here and here.

Face to Face interview

Advantages:

- Accurate screening. Face-to-face interviews help with more accurate screening. The individual being interviewed is unable to provide false information during screening questions such as gender, age, or race.
- Keep focus. The interviewer is the one that has control over the interview and can keep the interviewee focused and on track to completion.
- Capture emotions and behaviors. Face-to-face interviews can no doubt capture an interviewee's emotions and behaviors. Interviewer opinion can be a very good predictor of vulnerability for instance

Disadvantages:

- Cost. Cost is a major disadvantage for face-to-face interviews. They require a staff of people to conduct the interviews, which means there will be personnel costs.
- Quality of data by interviewer. The likelihood of the entire interviewing staff having those skills is low. Some interviewers may also have their own biases that could impact the way they input responses.
- Reluctance: Women and the elderly may feel more physically vulnerable than men and younger people. So the first of these groups may be more reluctant to allow a stranger into their homes for an interview, whereas they may be willing to talk with an interviewer over the telephone.
- Limit sample size. The size of the sample is limited to the size of your interviewing staff, the area in which the interviews are conducted, and the number of qualified respondents within that area.
- Higher level of "unknown" response are observed and were demonstrated by studies: "There is more item non-response in in-person interviews and that non-response is being driven by people with low levels of cognitive skills. The fact that there was an increased rate of correct responses to fact-based questions in the in-person interviews compared to the self-completed ones – with the concomitant decrease in "don't knows" – suggests that it is not the case that people are randomly guessing in the self-completed modes. The in-person interview seems to keep people from answering even when they know the correct response".

Telephone interview through Call Center

Telephone surveys may provide a good alternative, but we would advise use of a larger sample.

Advantages:

- Increasing rates of telephone coverage,
- Low cost of telephone surveys relative to face-to-face interviews,
- Speed with which telephone surveys can be conducted.

Disadvantages:

- People without telephones can only participate in face-to-face surveys. The systematic exclusion of this latter group from telephone surveys may introduce bias if this group is both sufficiently sizable and also sufficiently different from telephone owners.
- Young adults are more transient and less settled than middle or older age adults, the former may be less likely to own telephones and therefore may be under-represented in telephone samples.
- Telephone survey responses manifest more social desirability response bias than the Internet survey;
- Risk of selecting random people who lost/changed their phone.

Self administered with online quota survey

The emergence of Internet surveys in the 1990s threatened the dominance of telephone surveys due to their advantages in terms of cost and speed. Indeed, Internet surveys soon appeared as a promising alternative to prior methods; nevertheless, there are still problems with the coverage and, as a result, with the representativeness of online surveys.

Scrolling versus Paging, SMS versus E-mail Invitations have an impact on response rates in web surveys completed on personal computers. Scrolling design leads to significantly faster completion times, lower (though not significantly lower) breakoff rates, fewer technical problems, and higher subjective ratings of the questionnaire. SMS invitations are more effective than e-mail invitations in mobile web surveys.

Advantages:

- Lower cost and higher speed;
- Visual, interactive, and flexible;
- Do not require interviewers to be present and busy people – often educated and well-off – who systematically ignore taking part in a telephone survey are willing to answer questions posted on their computer screens;

- Studies found that Internet-based surveys increased the reporting on sensitive information, compared to computer-assisted telephone interview.

Disadvantages:

- under representation of uneducated population group: panel: one can reach only those who are online; one can reach only those who agree to become part of a panel; not all those who are invited respond; and, those who sign up for online panels are rather young and male.

Interactive Voice Response & Short text Message (SMS) Pools

IVR (Interactive Voice Response) can be used with a Free tool #.

Advantages:

- Accessible to illiterate and in multiple language
- Can support a lot of concurrent connection

Disadvantages:

- Technology need to be purchased
- Can not process to complex information

SMS gateway: FrontlineSMS has been already used under the project name Ascend and tested by UNHCR in Costa Rica

Specific constraints:

- 4 options max per questions
- concise questions, simple words
- describe action then press key
- caller should be informed on how much to go to complete the survey

Advantages:

- Lower cost and higher speed.

Disadvantages:

- only fit for short questions & short survey (less than five questions)
- Risk of selecting random people who lost/changed their phone

Interview incentives

A series of article about surveys on SGBV in US universities , here, and here, and here raised the point of the methodological approach to get SGBV statistics.

Low response rate may virtually guarantee exaggeration of certain participants.

For instance, incentives/fear make people declare information inaccurately and generate declaration bias. For instancem Fear associated with Legal obligations or Reporting Financial Issues in hope to get support

Interview length

The World Bank has studied the effect of the questionnaire length on survey results and found that identical questions asked of the same population yield different answers in short vs. long questionnaires due to differences in interactions between different questionnaire designs and respondent cognitive processes. This is visible through:

- variables, particularly those related to subjective welfare and housing, are impacted by changes in questionnaire design.
- households answer the same questions differently when interviewed with the short versus the long questionnaire during the same time period.
- differences in reporting are sufficient to yield poverty predictions that are significantly different in the short and long questionnaires.

Who to interview?

The selection of the respondent is absolutely critical especially when it comes to questions about perceptions. Men and women in the same household respond differently to individual questions. The fact that the degree of discordance varied so dramatically from one question to the next suggests that there is something about the questions themselves that affects how men and women will respond.

Different set of practice can be used:

- The household head is interviewed.
- Whomever is available when the enumerator shows up is interviewed.
- The member of the family who can speak the language of the interviewer. This is particularly the case in ethnic minority communities where men tend to speak better a vehicular language than women (since men generally have more relations with people outside the community and often have gone to school longer).
- Multiple members of the household are interviewed, with the most knowledgeable respondents providing different pieces of information wherever possible.
- A random adult among those present at the time of visit is interviewed
- A random adult among all those on the household roster is interviewed.
- Some questions are triggered for men and some questions are triggered for females only

Although it might not be the recommended approach, in practice the “Whomever is available” is usually

used. The design of the form can actually limit potential exclusion effects.

How to deal with Sensitive questions?

List randomization (also referred to as list experiments, the item-count technique, and the unmatched count technique) is a way of obtaining truthful responses to questions related to sensitive issues. This technique can be used for issues related to migration for instance (see this study or this reference page or this one).

The use of this method typically leads to higher reports of sensitive behaviors than is obtained through direct questioning. Practically speaking, individuals are randomly allocated into two groups. The two groups are offered different answers for the same questions - the different answer is the sensitive one. Subtracting the mean number of true statements reported by group B from the mean number of true statements reported by group A then gives the proportion of the sample that falls under the sensitive option.

Pre-Assessment

IMPORTANT

A few tasks need to be completed before the proper initiation of the assessment. These include:

- Informing key community stakeholders about the assessment,
- Gathering critical site information required for effective primary data collection,
- Confirming (or refuting) assumptions made in the planning stage, including logistics issues, travel time, limits of accessibility, and time available for fieldwork as well as highlighting implications of security issues (curfews, movements of armed actors, and risks of mines or explosive remnants of war).

It is recommended that, before actual data collection takes place in the field, a preparatory visit is done to each site of the assessment. This site preparation visit is the first contact between the assessment team and affected communities.

During site preparation (which can be undertaken by the field team leader, together with a second team member), meetings are held with key stakeholders in which team members will:

- Introduce themselves,
- Introduce the assessment objectives,
- Agree the assessment schedule and confirm that it is suitable to the community and that it will include representation of necessary population groups as planned, including those most vulnerable,
- Work with authorities, community leaders or members of the local community to identify KI interview and FGD participants,
- Confirm that the site has the characteristics of the strata it has been selected for,
- Identify locations for interviews and FGDs and confirm the suitability of these structures,
- Provide information about the work plan , confirm KI interview and FGD start times,
- Inform whether the participants should expect food or refreshments,
- Take the details of community leaders for any unforeseen communication.

A master list of sites which have been visited should be kept by the assessment team leader and noted on a map of the affected area to ensure that sites are not visited twice and that an appropriate geographic

selection of sites has been used.

Protection Topics

IMPORTANT

Translating assessment results into programmatic response recommendations is challenging when it comes to protection activities.

The analysis of vulnerabilities allows to generate protection intervention recommendations.

The analysis of vulnerabilities is complex as it should allow to understand:

- How **multiple vulnerabilities** interact between each other (prioritisation through criticality)?
- What is, for **each vulnerability type**, the specific profile of a population group in terms of magnitude and severity?

The Rights-based approach

Protection programme are designed according to the Rights-Based approach. As such, more than the needs, it is the vulnerability of specific population group towards rights violation that informs resources allocation.

Vulnerability level for each specific risk = function(risk **occurrence**) + function(**exposure** to risk of basic right violation) + function(**coping capacity** before or after violation)

Protection activities for UNHCR:

- Prevent occurrence of basic right violation event (**Environment Building actions**) &&
- Limit the effects of violation events consequences (**Remedial actions**) &&
- Respond to basic right violation event (**Responsive actions**)

Protection analysis is therefore linked to specific events linked to specific Right groups defined by UNHCR Result Based management.

Rights Group	Basic right type	Violation event
Favourable Protection Environment	Non-refoulement	Non-admission, Refoulement
-	No discrimination	Detention
-	Freedom of movement	Eviction, deportation
Fair Protection Processes and Documentation	Documentation	No issuance/ renewal of residency
-	Right to a nationality	Birth Registration
Community Empowerment and Self Reliance	Livelihood	Negative coping mechanism (Prostitution, child labour, early marriage, begging, stealing,.)
-	Peaceful co-existence	Intercommunity violence
Basic Needs and Essential Services	Education	Children out of school
-	Food	Malnutrition
-	Social & economic rights	Non access to services (healh, schools, MPSS)
-	Housing, Wash & Shelter	Sub standard living conditions
Security from Violence and Exploitation	Age, gender and diversity mainstreaming	Sex & Gender based violence, child abuses

Vulnerability Profile

On each topic, the protection assessment should allow to identify the affected population's vulnerability profile:

- Threat in terms of **criticality** , i.e. how important is the threat compared to other. This is measured through priority ranking.
- Threat in terms of occurrence in order to define **magnitude**
- Capacities and capabilities in order to define **severity** – "Severity" (or intensity) expresses the degree of unmet needs (it is thus related to shortages and deficits, as opposed to fulfillment and wellbeing) or the degree of something harmful,harsh, stern, irreversible or not desirable. As such it is expressed through a form of rating.

The vulnerability profile (combination of criticality/magnitude/severity) for each topic will allow to prioritise the relevant activities. Severity and Priority are therefore appropriate way to measure vulnerability.

The risk profile will help designing the best intervention approach for each group:

Activity	Criticality	Magnitude	Severity
Environment Building actions	++	+	-
Remedial actions	+	++	+
Responsive actions	+	-	++

Analysis Topics

Favourable Protection Environment

List of analysis modules

- Protection-sensitive border mechanism
- Counselling & Legal aid
- Freedom of movement & Detention
- Multiple Displacement & Movement

Fair Protection Processes and Documentation

List of analysis modules

- Birth Registration
- Civil status documentation
- Family re-unification
- Refugee registration
- Risk of eviction

Community Empowerment and Self Reliance

List of analysis modules:

- Social cohesion
- Housing, shelter & WASH (Water, Sanitation & Hygien) conditions
- Employment & livelihood
- Child Labour
- Assets & Budget

Basic Needs and Essential Services

List of analysis modules:

- Refugees with specific needs
- Assistance received
- Negative coping mechanism
- Food security
- Access to health services
- Education & Out of School Children

Security from Violence and Exploitation

List of analysis modules:

- Sexual Harassment & Violence (Rape)
- Harassment between youth
- Gender inequality
- Domestic Violence (Parental stress, Isolation at home)

- Forced and early marriage
- Violence among children
- Armed recruitment
- Harassment & Violence towards LGBTI
- Security & Civil violence

Module and questions

IMPORTANT

Objective of the standardised question modules:

- Avoid replicating **bad practices**;
- Ensure that all **topics of interest** are well covered in the analysis framework;
- Check that for each selected topics, all **required questions** to calculate the indicators and vulnerability profile (severity, magnitude, criticality) are included;
- Deliver **final reports** quickly through reproducible analysis workflow;
- Collect information at case level and record links between case that forms a household.

Intro: from the plan to the form

During the design stage of the assessment, the team must consider how each data element collected will be compiled, aggregated, analysed and disseminated to satisfy the information needs. The data analysis plans allow the assessment team to identify data elements on data collection forms that are not analysable as well as information management requirements that have not been met. Once missing data elements have been identified, forms can be amended accordingly by modifying, removing or adding data elements.

A common mistake for assessment teams is to collect too much data that is neither analysable nor will be used in decision making. The temptation to ask too many questions means that teams gather poorer quality data which obstructs useful analysis. If a data analysis plan shows that too much data will be collected, then data collection forms should be revised and shortened accordingly.

Each data element collected should be linked to:

- An information need linked to a Protection Topic
- Contextual information in relation with the Population Group (Refugees, IDPs)
- Data source - i.e. from aggregated Household Information, Key Informant or Focus Group discussions
- Type of Analysis: I.e. Correlation, Dispersion, Average,
- Specific Protection indicator

Unit of analysis: record at Case level and analyse household.

Household is the standard unit used in both national and international Household survey programme:

- Allow for comparability
- Allow for more complex analysis of interaction
- More complex to capture than case information

"Persons who live together and have communal arrangements concerning subsistence and other necessities of life, such as eating together" . As such Household includes two main concepts:

- The household dwelling -> living in a housing unit as belonging to the same household.
- The housekeeping concept -> common provision for food or other essentials for living, with or without combining with any other person to form part of a multi-person household.

Household can corresponds to the 3 follwoing differents types:
* Nuclear household (Family unit)

* Extended household (Family unit + additional related members) * Composite household (Family unit + additional non-related members)

Case (i.e. group of individuals as registered) corresponds to the unit used for assistance, as registered by UNHCR and also used during continuous assessment / protection monitoring. UNHCR Registration process is described in the Registration handbook

It is recommended to have one form filled by each case rather than by each household. In this form one repeat questions can allow to records all linkeds cases with their respective CaseID for the following reasons:

- It's easier, after data collection, to reconstruct a household from multiple case than deconstruct household in multiple cases as this implies to define the correct elements to re-apportion correctly the right information.
- The types of links between cases (level of sharing for accomodation and for food) that forms an household and the household profile can easily be collected by case and then triangulated between cases.
- There's a risk to miss some extremely vulnerable cases if their record is merge with a less vulnerable, creating either an exclusion error or worst generating some 'Cobra effect'.

Questions modules

The Household level Assessment includes different parts that includes a series of modules. The selection of required modules will be done through a participatory assessment. Each module allows to define a series of indicators. Questions to be used in each module are extracted whenever possible from the IHSN Question Bank.

Introduction

Introduction should include the follwoing mains elements

Enumerator

Consent

Part 1: Background Information

Eligibility: To be administered to every household in the main sample.

This questionnaire serves three purposes: (i) to identify the members of the household; (ii) within households, to identify nuclear units, i.e. couples and their own children; (iii) to collect basic demographic information on each of the household members;

- PA status
- Household information
- Address
- Refugee registration
- Reason for displacement
- Multiple Displacement & Movement

Part 2: Household condition

Eligibility: To be administered to every household in the main sample.

- Housing, shelter & wash conditions
- Assets & Budget
- Food security & Negative coping mechanism

Part 3: Household composition

Eligibility: To be administered to every member of the household.

- Employment & livelihood
- Education & Out of School Children & Child Labour

Part 4: Basic Needs and Essential Services

Eligibility: To be administered to every household in the main sample.

- Access to health services
- Assistance received - Post Distribution Monitoring
- Self-expressed Needs (Hesper scale)

Part 5: Protection & Rights

Eligibility: To be administered to every household in the main sample.

- Civil status documentation
- Birth Registration
- Counselling & Legal aid
- Freedom of movement & Detention
- Risk of eviction
- Refugees with specific needs

Part 6: Attitude & perception

Eligibility: To be administered to one male member and one female member of the household.

- Return Intentions
- Onward movement
- Armed recruitment
- Domestic Violence (Parental stress, Isolation at home)
- Forced and early marriage
- Gender inequality
- Harassment & Violence towards LGBTI
- Harassment between youth
- Security & Civil violence
- Sexual Harassment & Violence (Rape)
- Violence among children

Conclusion

Additional elements to be filled by the enumerator at the end of the interview.

- Subjective Enumerator Evaluation – clearly outlining which elements should be considered for each vulnerability category
- Referral (if needed)

Question design

Checklist used to review questions

- Closing questions
- Skip patterns
- Ranges for selected questions
- Numbering questions
- Two questions in one
- Double negatives
- Clarity and simplicity
- Consultation
- Parcimony

Context information

- Metadata
- Enumerator
- Address
- Consent
- Household information
- Self-expressed Needs
- Subjective Evaluation
- Referral
- Notes

Locations & Geography

Pcoding location within a form is an important taks

It is possible to add pcode through a reference file, *itemsets.csv*, which is available for each country. This reference file can be used when creating Cascading selects questions using the select_one_external function within XLSFORM questionnaires. The itemsets.csv file can be uploaded to UNHCR Kobo Server as a media file. It will be downloaded to any Kobo Collect like any other media file and saved to the [form-filename]-media folder. Clients like Kobo Collect load media files from the SD card and so a field data collection form with all locations within the country can load very quickly.

itemsets.csv for the MENA region countries can be downloaded from UNHCR MENA pcode

Likert questions

Likert scale - also called "opinion scale" - describe a way to formulate questions related to opinion. A Likert item is simply a statement that the respondent is asked to evaluate by giving it a quantitative value on any kind of subjective or objective dimension, with level of agreement/disagreement being the dimension most commonly used.

Rating questions are often much easier to understand by individual than **ranking questions** where people may be forced to make one item worse or better than another, when they actually find them equal. Psychological research has proven that it's a very tough thing for people to rank more than three options. When people are given a cognitively difficult task like ranking question choices of for instance six things, often they will end up getting frustrated and will start to rank choices randomly. This will lead to higher dropout rates and even worse, messy data.

Likert questions allows for each individual part to have the same value, if that is how people feel about

them. Last the analysis of a rating is more intuitive than the one of a rank that in addition might not have much statistical significance. Also from Likert questions, it's possible to compute an average score from each separate ratings in order to get an overall ranking through a Borda Count.

When designing Likert questions, attention need to be given to specific design issues specifically in terms of wording.

1. Length of the scale: limit the number of points along the scale.
2. Odd or even point scale: There should not be preferred or better choice.
3. Label the points in the scale: Avoid using just numbers to indicate the points on the scale.
4. Balanced scale: Make sure that the scale is balanced with an equal number of positive and negative categories. Center point on bipolar scale: A common mistake when creating a rating scale is including "no opinion" or "uncertain" as a middle response on a bipolar scale. These options are not actually a part of the scale order. A middle category in a scale between "agree" and "disagree" would be "neither agree nor disagree". Options such as: "no opinion", "not sure", "undecided", "don't know", or "not applicable" are placed off the scale, in a separate space.
5. Match response to question: Be as direct and specific as possible, focusing the response options on what you want to measure.
6. Keep labels consistent: Finally, the labels used in the scale need to refer to the same thing.

Some examples are given below (more examples here):

Agreement

- Strongly Agree
- Agree
- Undecided
- Disagree
- Strongly Disagree

Frequency

- Very Frequently
- Frequently
- Occasionally
- Rarely
- Never

Importance

- Very Important
- Important
- Moderately Important
- Of Little Importance
- Unimportant

Likelihood

- Almost Always True
- Usually True
- Occasionally True
- Usually Not True
- Almost Never True

Quality

- Very poor
- Poor
- Fair
- Good
- Very Good

Interest

- Not at all interesting
- Slightly interesting
- Moderately interesting
- Very interesting
- Extremely interesting

Using the modules

The XLSFORM format

XLSFORM is a standard way to describe a form / questionnaire. It includes information not only on the type of questions but also on:

- Hints in order to add comments on methodology
- Contraints to avoid entering incorrect information
- Relevant to allow for skip logic
- Support for multiple language
- Possibility to perfom calculation for data quality checking
- Possibility to repeat the same questions in a loop.

The next step will be then to customise the form, page 57

Questionnaire analysis report

The libary can generate a short report that summarises:

- What topics are covered by the form
- Check that for each topics, questions are selected to allow for the estimation of magnitude, criticality & severity
- Summarise questions that can be raised in key Informant Interview, samples household interview, on site household interview or Focus group Discussions.

Guidelines for questionnaire customisation

NOTE

Introduction

A great deal of effort has been devoted to ensure that the questions can be used by interviewers to obtain answers from respondents that are both reliable and valid. Thus the questions are drafted in a clear simple language and follow clearly and logically from one to the other, while the layout is designed to make it easy for interviewers to administer the questionnaires. The wording and question sequence are designed to motivate respondents and help them recall information on past events. Using the model questionnaires verbatim is most likely to ensure that the results of the surveys are comparable across participating countries.

It is therefore important that even if adaption and contextualisation are made, the wording of the model questionnaires is not drastically changed.

What is “customisation”?

Customisation (or adaptation) refers to the process during which the proposed Protection Assessment questionnaires are tailored to the population/context where a the Assessment is being conducted (that is, a national assessment, or a assessment conducted for a population group or for a selected area within a country), using standard principles and approaches, while maintaining global comparability of the indicators that will be derived from the collected data.

The customisation process is by no means an easy and straightforward one. Without a detailed understanding of all the standard tools and of the general principles and recommendations, customisation of questionnaires should not be attempted at country level without the assistance of an expert. During the customisation process, it is also absolutely critical that lessons learned from previous data collection activities are used effectively, and wherever necessary, tools are tested before final decisions are made. Testing may include organized pre-testing, field testing, piloting, and in some cases, cognitive testing. Analysis of raw data from previous assessment and data collection activities, as well as results from these efforts should also be undertaken for successful customisation of standard questionnaires.

What kind of “Customisation”?

Customisation of the “Questionnaire Modules”, “Questions” and “Response Categories” are necessary for at least two basic reasons: * No single country/survey is expected or recommended to use all of the modules in proposed questionnaires * No single standard questionnaire can accurately represent all human experience around the globe

Customisation covers the following types of changes to the standard protection questionnaires: * Country-/assessment-specific modifications to already existing standard questions and response codes, * Deletions from the standard questionnaires, and * Additions to the proposed questionnaires.

Modification

Certain parts of the proposed questionnaires must be modified. Indeed, in several instances, the proposed questionnaires include clear directives that a change or modification needs to be made. These cases are indicated using text such as “insert local name”. Similarly, response categories that require customisation are also indicated.

Deletion

No assessment is recommended to retain all of the modules and questions of the proposed protection questionnaires. First, there will always be some topics that will not be relevant in certain countries or regions. Second, decisions on the content of any assessment will ideally be made as a result of a thorough data gap assessment, generally based on the required analysis, and, for example, when information is available from other recent data sources, certain modules or sets of questions will be dropped. The process and analysis involved in a comprehensive data needs assessment will vary, but is a crucial step in determining the content of the assessment.

Determining what to exclude from the assessment is a balancing act that should take data needs into account, but also learn from countless experiences of data quality issues as a result of overloaded questionnaires. Country priorities will guide decisions, but may also work against achieving an optimum questionnaire size if negotiations turn more political than technical.

A final consideration will also rest with the ability to implement an adequate sample size, as this is often constrained by budget on one hand and on the other the known data quality issues associated with large sample sizes. For instance, some indicators are difficult to measure in low fertility settings, demanding higher sample sizes or complicated sample designs. Unless such issues can be technically addressed, the exclusion of such indicators may be necessary.

Addition

Some Protection Asessment may also add topics, modules and questions which are not already in the proposed questionnaires. These could include additions that the proposed questionnaires already point to (for example, adding household assets to the list already in the questionnaires), or additions of modules or sets of questions that are not covered in the proposed questionnaires.

From the onset of considerations of what could be added in, you should know that this will affect the technical support available as well as require changes and considerations throughout the package of tools available, from sampling, training, instructions, and data entry application to tabulations and reporting.

As with the above exercise of deleting from the questionnaires, your entry point should be the indicator list or, alternatively, the tabulation plan. Questionnaire design is secondary to the need for precise information on what such proposed additions would be measuring and how such would be presented.

Only questions that are previously well-tested and validated should be included. Questions are often imported from other household surveys that have been conducted in the country. This does not necessarily mean that they are validated nor does it mean that such questions can work within the frame of a Protection Assessment.

If additions are made, please ensure that formatting and coding follow the rules in place for the proposed

questionnaires. For entirely new topics it may be useful to build a new module and in other cases you will need to append to an existing module or insert within the existing flow. If you create new questions, submit them for addition in the library here.

Rules and Useful Tips for Customisation

Customise but do not compromise global comparability: * Assess the implications of changes; * Check that all required questions to calculate the indicators are included; * Check previous surveys to see how the customisation was done; * Consider translations for all major languages spoken among the survey population – Arabic, Kurdish,etc..

Pre-test rigorously to make sure that : * The questions are understood and the response categories are meaningful; * The language style that can be understood by everyone; * The skip-logic functions within the form are working well;

Translation

Translation should be planned as an integral part of the study design.

Importance of translation

Mistaken translation can greatly jeopardize research findings. As reported in the article World values lost in translation, many translated terms showed different associations than the term used in English. It also shows the changes of translation in later waves of the survey made trend analysis impossible. It thus prevents the analysis on the stability of change in values, which is one of the main goal of many survey.

Translation costs will make up a very small part of a survey budget and cannot reasonably be looked at as a place to cut costs. Experience gained in organizing translation projects and selecting strong translators and other experts is likely to streamline even these costs. A professional translator does 300 words/hour, **2500 words/day** and a 40 minutes long questionnaire can be around 4000 characters long meaning that the initial translation could theoritically be done in 2 days. Translation cost for freelancer can vary but a benchmark of 10 cents/word can be used (meaning that the outsourced translation of a 4000 characters long questionnaire would cost about 400 USD).

Avoid unwritten translation

When the same questionnaire is runned to different population group from different linguistic background, it may become necessary to have bilingual interviewers as they will have to present and read the questions in multiple language.

Sometimes, in addition of presenting and reading in multiple language, bilingual interviewers translate for respondents as they conduct the interview acting as interpreters. In other words, there is a written source questionnaire that the interviewers look at but there is never a written translation, only what they produce orally on the spot. This is sometimes called "on sight", "on the fly" or "oral"" translation.

Evidence available from recent investigations suggests that these modes of translation must be avoided whenever possible for the following reasons:

- There will be inevitabily some variance in translation performance resulting in potential misinterpretation of results.
- What might be gained in saving from translator will be lost with the extensive training and briefing that will need take place in order to disambiguate all complex terms and concept in the survey.

Organising translation process

Translation is an important step when finalising the survey. The important elements to considerer are summarised as TRAPD (Translation, Review, Adjudication, Pretesting, and Documentation):

- **Translators** produce, independently from each other, initial translations, (they provide the draft materials for the first discussion). Each translator may prepare a full translation (double/parallel translation) or the material to be translated may be divided among the translators (split translation). Review of the first 10% of the initial translation (in case you are working with a new translator) may indicate that a given translator is not suitable for the project because it is unlikely that serious deficiencies in translation quality can be remedied by more training or improved instructions. If this is the case, it is probably better to start over with a new translator.
- **Reviewers** review translations with the translators. Translators should not simply "hand over" the finished assignment abut should be included in the review discussion. Note relying on one person to provide the initial questionnaire translation is particularly problematic if the review is also undertaken by individuals rather than a team (these are reasons for working in teams rather than working with individuals). Even if only one translator can be hired, one or more persons with strong bilingual skills could be involved in the review process. Part of the review is to check for general tone consistency: this means that it is important to use the same style in the entire survey instrument, in terms of language register, politeness norms or level of difficulty. Some projects rely on procedures variously called "back translation" to check that their survey translations are adequate: the two source language versions are compared to try to find out if there are problems in the target language text. In practical and theoretical terms, it is recommended to focus attention on first producing the best possible translation and then directly evaluating the translation produced in the target language, rather than indirectly through a back translation. Comparisons of an original source text and a back-translated source text provide only limited and potentially misleading insight into the quality of the target language.

- One (or more) **adjudicator** decides whether the translation is ready to move to detailed pretesting and also decides when the translation can be considered to be finalized and ready for fielding. Official approval emphasizes the importance of this step and the significance of translation procedures in the project.
- **Documentation** of each step is used as a quality assurance and monitoring tool. It is also important if the survey has to be reproduced.

Reference

To know more about the subject, one can consult this Bibliography on the transaltion of survey, some important reference includes the Lessons from the European Social Survey (ESS), the Census Bureau Guideline for the Translation of Data Collection Instruments and Supporting Materials or this Procedure to prevent differences in translated survey items using Survey Quality Prediction program.

Configure forms

IMPORTANT

Configure forms can be done both through the Kobobuilder interface or within Excel. using kobobuilder comes with Risk when the form is complex.

The second options is preferred as it allows for more controls.

XLSFORM

The form needs to be defined using the xlsform format. XLSFORM is a form standard created to help simplify the authoring of forms in Excel. Authoring is done in a human readable format using a familiar tool that almost everyone knows - Excel. XLSFORMs provide a practical standard for sharing and collaborating on authoring forms. They are simple to get started with but allow for the authoring of complex forms by someone familiar with the syntax described below.

The XLSFORM is then converted to an X-Form, a popular open form standard, that allows you to author a form with complex functionality like skip logic in a consistent way across a number of web and mobile data collection platforms.

The documentation on xlsform is here. The same format is actually used by multiple data collection technological platform such as kobotoolbox(the platform used by UNHCR on http://kobo.unhcr.org), OpenDataKit or ONA.

Some tips

Good questionnaire design can save hours at latter stage for the person that will need to analyse the data…. do invest enough time for this step and allow for peer review before starting the testing phase !!

Manage your xls form

- Make sure your file is saved in the .xls format and contains no spaces or special characters ('-' and '_' are allowed).
- Make sure that your column headers are in lowercase (i.e. "label" or "name", not "Label" or "Name")
- Make sure that your sheet names are appropriately named (i.e. "survey" not "Sheet 1", "Survey" or "surveys")
- Make sure that the question names are unique and do not contain spaces or special characters ('-' and '_' are allowed).
- By the end of the coding sessions, there will be several versions of the MS Xls files. It is crucial to separate each according to versions and also to make sure the names do not in clude any spaces or begin with numbers or special characters.

Variable Encoding

- The "name" column must be within 32 characters, cannot begin with a number or any special character and cannot contain any spaces in between; otherwise ODK Validate will issue an error. The longer your name is and the more time you will have to spend searching for what that specific variable or choice is in your database. Also, Kobo doesn't accept answer choice names that contain more than 32 characters and question/ variable names no longer than 30 characters
- Please use a short meaningfull name for your variable rather than a code like Q1 or Q2. The name should mean something to you and should be clear as to what it refers to.
- Having a consistent and harmonised naming system throughout your survey, with descriptive/ meaningful and easy to remember names for your variables/questions and choices, will allow you to quickly find the variable(s) to be modified. On the contrary, should you have a non-harmonised naming system with names that are non-descriptive, it will take you much more time

than expected to find the variables of interest and to modify them.

- In order to anticipate the analysis with a statistical package like koboloadeR, limit variable names lenght to 12 characters and do not start a variable name with a number.
- Avoid underscores («_») in the technical name of the variable. It can sometimes be recognised as a «separator» by certain statistical packages. Do not hesitate to rather use the «camelCase» (consists on writing a group of words in lowercase with the first letter of the lined words in uppercase).
- Name should not contain any spaces, accentuation ("à è ô") or special characters ("+", "%", "&","(", "/", etc.).
- Names should not start with a number: You can use numbers in your names, but rather at the end or in the middle of your name so that it is instantly clearer as to what information it refers to.

More tips on variable name encoding are available here

Labeling questions

- Numbering questions does not provide added value – during data entry enumerator are forced to respect the order any way. In addition, if you start numbering questions, any change in the questionnaire will then require to renumber all questions. It's a good example of something that was important when people were using paper form but has no reason to be used anymore when moving to digital data collection.
- Quick translation can be done with Google translate. Obviously the resulting translation needs to be carefully reviewed but it save the time to retype the whole text and does not prevent from the translation review described here. It's also a good way to test the simplicity of the sentence used for the questions: if the artificial intelligence from Google translate can not decipher the questions, it means that it will be also probably difficult for the respondant. One test to confirm the question phrasing can also to translate and re-translate back in the original language in order to see if the meaning of the question is preserved.
- When there's a question like *first reason*, *second reason*, *third reason* or *first source of income*, *second soure of income*, *third source of income*, the bad practise is actually to have 3 separate select_one questions: aka *first reason* - then *second reason*, then *thirst reason* – First, from a cognitive point of view, respondents will have a hard time to rank among more than 5 options, second when recorded through 3 distinct variables, data will need a lot of coding to reshape the data in way where the analysis will make sense… – here the 2 options are either:
    - select_multiple with a constraint count on the number of modality you can select "`not(count-selected(.) < 4`" — often it's simpler as very likely the ranking between options will be difficult to analyse
    - or go for a rating likert questions for each options – and then the ranking will be

performed during the analysis through the sum of the rates for each respondent.

Questionnaire structure

- All “begin group” types must have a corresponding “end group”. It is advised to name each of the begin group and end group “name” as well as “label” columns for easy identification of where a group starts and ends, e.g. "type– begingroup. Name – grp_agri_aman. Label – Aman Rice. T ype – end group. Name – endgrp_agri_aman. Label – no labels are required as this won’t show up in the form.
- There can be a mother group with several child groups within, however each group essentially works like brackets which open, can remain open, close or close at the end. It can get rather complicated once several questions within a group are to be skipped.
- DO NOT LEAVE A SPACE AFTER your “type” or “name” variables, e.g. “end group”. It is an invisible error that will cause the form to malfunction.
- It is a good practice to keep naming the questions “label” as “Name of Household Head… Age of Respondent…Please specify what you mean by”Other“…and so on”. The same can be done for “name” variables but one must remember that they cannot start with an uppercase letter, e.g. “name_hhh” is allowed.

Make your questionnaire smart

- Use hints as much as possible - a good example is to recall to your enumerator how to proceed with select question for instance two options are possible :
    - Read the list and select all (or only the right) answers
    - Do not read the list - wait for the answer to select the right answer(s)
- The “note” and “calculate” type of questions can be really handy and must be used whenever there is a reason to check back with data entered previously. E.g. “type – calculate, name – calc_loan, calculation - ${loan_1}+{loan_2}+{loan_3}” with "type – note, name – chk_total_loan , label – Your breakdown of total loans ${loan_1} and ${loan_2} and ${loan_3} add up to ${calc_loan}. Please make sure they are equal to the amount entered before as total loan ${totalloan}.
- Systematically add constraint to your numeric questions: age can not be negative or bigger than 120. Add also constraint on linked values: for instance if you know the number of children in a family, the number of boys can not be bigger than the total number of children. Do not forget to add **constraint_message** to help your enumerator to fix the issue!
- When you do a calculation, **beware of zero** : To convert empty values to zero, use either the coalesce() function or the if() function. See the form logic functions reference below. "+ - *" correspond to add, subtract, and multiply. Division however, is special, and you need to use the

word “div” to do division.

- The best types of constraint and relevance are usually the simplest. When typing the binding guidelines (http://opendatakit.org/help/form - design/binding/) can simplify the entire line of code. Also this is the most interesting and perhaps challenging part of writing questionnaires. Users are free to develop their own lines of code and put them to the test. For any help the Google ODK groups are more than sufficient.
- If the questions need to be translated into a local language, the input must be in **Unicode**. Also it is imperative to make sure that questions in the base language are translated with the meaning intact. Difference in interpretation can cause severe errors in the quality of the data.
- Systematically add the metadata field: type: ‘start’, ‘end’, ‘today’, ‘deviceid’, ‘imei’, ‘phonenumber’
- To record **admin level**, use cascading drop down lists. Note that this does not work well whem more than 500 options. In this case, one option is to use select_external : http://xlsform.org/en/#external-selects but this will work only if data collection is done with Android client (not the web client).
- Record correctly registration case number using a regex constraint where caret is the ASCII special character:
    - for case number: “ `regex(.,‘caret[0-9]{3}-[0-9]{2}[C][0-9]{5}$’)` ”
    - and for Individual number “ `regex(.,‘caret[0-9]{3}-[0-9]{2}[I][0-9]{5}$’)` ”
- Record case number using bar code scanner whenever possible - rather than manual data entry.
- Add constraint on email validation through the same regex: “ `regex(., ‘[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+.[A-Za-z]{2,4}’)` ”. This can be more specific if a specific email domain is expected.
- Calculate years from date (age from date of birth) “ `round((decimal-date-time(`*today*`) − `*decimal* `− `*date* `− `*time*`({dob})) div 365.24, 0)` ”.
- Validate past date (constraint) “ `decimal-date-time(.)<decimal-date-time(${today})` ”.
- Use “or_other” whenever you have a select_one or select_multiple questions that may have additional modalities that you can not anticipate
- If you have a select_multiple question with one modalites being “None of the options mentionned here”, use a contraint to ensure that respondant will not select this options with other: “ `not(count-selected(.) > 1 and selected(., ‘no’))` ”
- Review the **appearance** options : Beware some options (like) are only for web client while other (like “quick”) are only for the android client.

Keep your xlsform legible

- It is possible to separate different sections of the questionnaire using different cell colours. It

does not make a difference for ODK, but makes the form much more navigable.

- In the same idea, in Excel, freeze the top row cell and add automatic filters to quickly navigate through your questions (survey) and answers (choices).
- In the choices sheet, a gap or (blank row) can be left after each set of answers. Again, it doe s not make a difference for ODK but makes it much easier for navigation.

Validate & Deploy the form

- kobotoolbox has some built-in **validation check & debugging check** when a xlsform is uploaded. Sometimes better validation / debugging support can be obtained by other online validator such as https://opendatakit.org/xlsform/
- Note for field deployment: if possible, users must try and keep a copy/image of the entire " ODK " folder inside the memory of the tablet and save it elsewhere for later reference. If planning requires users to clear out " ODK " and re insert a questionnaire for easy tracking, then this step becomes very a vital safe guard.
- one point is also that as for dataviz, good questionnaire are the result of multiple persons looking at it – hence the importance of peer review – peer review is different from testing - it requires people with expertise in questionnaire design to review - this is among the support than can be requestioned from regional offices.

Additional tips

Additional tutorials and tips can be found through the following links:

- Open data kit Form Logic
- Open data kit Form operator functions
- Open data kit Form design guides
- Open data kit work around
- Other training guides
- Tips from the Mobile Data Collection Toolkit
- Advanced XLSFORM coding tips #1
- Advanced XLSFORM coding tips #2
- ACF XLSFORM coding tips

Support Forum

When faced with specific challenges, multiple online forums can be consulted (as xlsform is used on different platform):

- Forum for Kobotoolbox
- Forum for OpenDataKit

Pre Test Phase

What is and why “pre-testing”?

Testing materials before they are used in live surveys allows to make certain that questionnaire is accurate, non-leading and reliable. This is often carried out at a late stage and using finished materials, i.e. the final version of a the questionnaire.

Considerable professional and financial investment is therefore at stake, and pre-tests can constitute some of the most difficult and contentious step within a household survey project. People often think that testing a survey takes a long time. They think they don’t have the time or resources for it, and so they end up just running the survey without any testing. This is a big mistake. Even testing with one person is better than no testing at all. So if you don’t have the time or resources to do everything in this guide, just do as much as you can with what you have available.

As a good practice, the entire questionnaire should be submitted to a number of procedures to ensure there are no questions likely to test anything but respondent real status or opinion. There are extensive guidelines on this subject, belwo is a summary of the main points to look at.

Objectives

The overall objectives and the focus of the pre-test, together with the results shoudl be properly documented before starting it.

- Questions are clear i.e. respondents do not **misinterpret** the questions (questions are not ambiguous or difficult to understand);

- **Response categories** are comprehensive and adequate for the assessed population; any answer falling into the "other (specify)" category of a multiple choice question and that constitutes about 5 percent or more of all answers to that question should be considered as a serious candidate for a separate answer category of its own. New codes for common answers that were not included in the original questionnaires will need to be created
- **Flow of questions** within the questionnaires is adequate;
- Difficult or **sensitive questions/modules** are identified so that extra training can focus on these questions during the fieldworker training;
- Translations are accurate, ie. **translated questionnaires** are working correctly. Changes in wording or improved translation will need to be incorporated when required;
- Interviewer **instructions** in the Instructions for Interviewers, as well respondent **informed consent** in questionnaire are clear and sufficient;
- Average **duration of interviews** is calculated in order to plan the fieldwork and the daily workload per interviewer/team

Document the organisation of the Pre-test

- **Clusters/Number of interviews selected for pre-test**: Describe the pre-test locations where were the households located for the pre-test, how and why were these locations selected, etc. Note that often, what is important during the pre-test is more the quality of the observation within each interview than the number of interviews itself.
- **Personnel**: Present the trainers and interviewers (trainees) of the pre-test. Include information on the future involvement of the participants in the rest of the assessment process.
- **Training**: List the dates and content of the pre-test training, as well as how it was organised. Some detail is useful on agenda and training methodology, as it can serve as lessons for the main training. Include other details as relevant: Venue, recommendations for main training, etc.
- **Fieldwork**: Provide the dates of actual pre-test fieldwork. Also detail on organisation (logistics, teams, areas, etc.) is very useful.
- **Findings**: Describe how the observations from the pre-test were collected and discussed and what process for making changes to the final questionnaires was used.

How many respondents?

As a general rule, one should aim to pretest a form with at least 5 to 10 respondents and 2 different interviewers. If the survey is more complex, as studied by reasearcher, a **sample of 30 respondents is**

recommended.

Even with this small number of people, a surprisingly large number of improvements can be made. Try to get within those 5 to 10 respondents a range of different people who are representative of the target group for the questionnaire. Usually, most of respondents will have the same problems with the survey, so even with such small number of people, it should be possible to identify most of the major issues. Adding more people might identify some additional smaller issues, but it also makes pretesting more time consuming and costly.

How to observe pre-test?

Note that for the pre-test, it's important to have 2 persons supporting the interview. One regular enumerator and an additional person to take note of each observation during the interview.

- Respondents should be asked to complete the survey while **"thinking out loud"**. They should tell you exactly what comes into their mind so that the observer can take notes on everything they say. This which can include paraphrasing, providing retrospective thinking or providing judgments of their confidence in what each question means.
- The observer should look for places where the respondent **hesitate or make mistakes**.
- A **debriefing** can be organised after each pre-test interview to ensure that potential additional observations from the enumerator are also inclued in the observation notes.

Pre-test Results and Recommendations

Relating to the objectives listed above, this section should include findings from the actual data gathered as well as the qualitative findings from the pre-test, including those obtained from discussions with interviewers after the pre-test fieldwork concluded.

- **Questionnaire:**
  This section is the main output of the Pre-test Report. The use of the table below is recommended. Please add all modules in the questionnaire. Make sure that all suggested changes are listed and that evidence is provided for final decisions. Please include observations on all country-specific modules and questions.
- **Instructions:**
  Describe and list any changes or additions required in the Instructions for Interviewers as well as those introduced in the Instructions for Supervisors and Editors. Such changes typically involve translation issues, instructions for country-specific questions, but also for country-specific response categories. Appropriate corrections are incredibly helpful and will especially inform the main field work training. Note that instruction can be introduced as additionnal *hint* within each question.

- **Average duration of interviews:**
  Calculate the average duration of interview for each questionnaire using the data collected in the pre-test. Typically, as interviewers become more familiar with the tools, this time will decrease and therefore a realistic duration should be proposed and included in the introductory sentences on the cover pages of the questionnaires.

- **Interview process considerations:**
  Describe and address the observations from the pre-test that relate to interviewing that will be relevant for training and monitoring in the main field work (for example issues in approaching households, dealing with sensitive module and questions, flow of field work, roles and responsibilities, etc.)

- **Assessment process considerations:**
  Describe here the observations, suggestions, and decisions related to the assessment planning and next steps for finalising the questionnaire (training contents/agenda, logistics, staff, support, etc.)

Fieldwork Training Agenda

Training

Training of survey teams is fundamental for the quality of the survey information collected and preparation for the training and survey will take time, especially the first year the full assessment is implemented.

The training for the full assessment is recommended to last at least 2 days, depending on context and team experience. Extra staff should be trained in case someone is unable to perform the field work.

The main topics to cover in training of data collectors (note that team leader may be provided with a more in-depth training than some of the data collectors) are as follows:

Day 1

Morning

- Reason / objectives for Protection Assessment
- Composition of survey teams : roles and responsibility
- Sampling procedure: why sample?, explained in a way that surveyors can later on explain to community members when asked; and rationale and importance of representativeness
- Questionnaire and sheets: household-level information, child-level information, woman-level information, observations

Afternoon

- Introduction to the household and informed verbal consent
- Interview questions and interviewing techniques: go through each question for clarity, answer options, cultural appropriateness, gender sensitivities, avoid suggestive questioning but probe where necessary
- Age recording and use of local events calendar
- Practicing with real children and/or adults

Day 2

Morning

- Survey logistics
- Equipment
- Communication
- Travel
- Incentives / salary/allowances
- Food and drinks
- Accommodation, etc.

Afternoon

- Pilot test

Data Protection Impact Assessment

IMPORTANT

This chapter is not written yet.

UNHCR Data Protection Policy

The Policy on the Protection of Personal Data of Persons of Concern to UNHCR has been issued in May 2015.

Data Protection Policy

In addition of the clarification of the basic principles to be applied when collecting information that

includes **Personally Identifiable Information (PII)**, the policy recall the needs to carry out **Data Protection Impact Assessment (DPIA)**. A DPIA is required where the collection and processing or transfer of personal data is likely to be large, repeated or structural.

Bceause most of "Protection Assessment" fall under that policy as they are linked with the need to be able to process referals, personally identifiable information are often required. In addition, information such a precise coordinates of individuals with specific profile or needs might also fall unde the policy.

The 12 privacy principles

Data Protection Impact Assessments (PIAs) are an integral part of taking the "privacy by design" approach. This is done by enforcing the following 12 principles:

1. **Consent and choice**: presenting to the data subject the choice whether to allow the processing of their personally identifiable information (PII)
2. **Purpose legitimacy and specification**: ensuring that the purpose of data collection is specified and lawful
3. **Collection limitation**: limiting the collection of PII to that which is within applicable law and strictly necessary for the specified purpose(s)
4. **Data minimisation**: minimising the PII processed and the number of privacy stakeholders to whom PII is disclosed or who have access to it
5. **Retention and deletion**: ensuring that data is not kept for longer than is necessary for the purpose specified
6. **Accuracy and quality**: ensuring that the PII processed is accurate, complete, up to date (unless there is a legitimate basis for keeping outdated data), adequate and relevant for the purpose of use
7. **Openness, transparency and notice**: providing data subjects with clear and easily accessible information about the PII controller's policies, procedures and practices with respect to processing of PII
8. **Individual participation and access**: giving data subjects the ability to access and review their PII, provided their identity is first authenticated (Access and correction)
9. **Accountability**: assigning to a specified individual within the organisation the task of implementing the privacy-related policies, procedures and practices
10. **Information security**: protecting PII under an organisation's control with appropriate controls at the operational, functional and strategic level to ensure the integrity, confidentiality and availability of the PII, and to protect it against risks such as unauthorised access, destruction, use, modification, disclosure or loss.

11. **Privacy compliance**: verifying and demonstrating that the processing of data meets data protection and privacy legislation by periodically conducting audits using internal or trusted third-parties
12. **Data transfers**: do not store or transfer personal data to third parties without adequate assurances that they will safeguard it to a standard comparable to that of the UNHCR.

Threat & vulnerability matrix

The template from the UK ICO can be used a starting point to developd the document.

An important point is to build the threat & vulnerability matrix:

Potential Threats to look at	Vulnerability	Risk	Mitigation
Cyber espionage			
Physical loss of data			
Technical failure			
Unauthorised acquisition			
DDOS attack / malware			
Insider privilege abuse			
Partner abuse			
Partner negligence			
Refugee complaints litigation			
Reputational damage			

Deleting dataset

Personal data that is not recorded in individual case files is not to be retained longer than necessary for the purpose(s) for which it was collected.

Though this should not minimise the need to save anonymised copies of the dataset in order to ensure potential new analysis of the data or longtidinal analysis. See the last chapter for more information on this, page 153.

Technical measures necessary to comply with the policy

- Maintaining physical security of premises, portable equipment, individual case files and records;
- Maintaining computer and information technology (IT) security, for example, access control (e.g. passwords, tiered access), user control, storage control, input control, communication and transport control (e.g., encryption).

Standard Operating Procedures necessary to comply with the policy

A DPIA would contain a general description of:

- the envisaged project,
- data sharing agreement in place
- arrangement involving processing of personal data (for instance clarififcation on protection referal),
- analysis of the risks to the rights of data subjects by virtue of the circumstances
- nature of the personal data processed,
- safeguards and security measures in place or proposed.

KoboToolBox Server Configuration

IMPORTANT

This chapter is not written yet.

What is KoboToolBox?

KoBo Toolbox is an open-source tool for mobile data collection. It allows you to collect data in the field using mobile devices such as mobile phones or tablets, as well as with paper or computers.

The project has been orginally developipped by the Harvard Humanitarian Initiative and was then adapted to humanitarian usage in colleaboration with OCHA.

KoBoToolbox was first created in 2009 and has grown over the years to include several projects and initiatives. It includes:

- **FORMBUILDER**: Easily create survey forms through an online user interface.
- **MOBILE DATA COLLECTION**: Quickly and reliably collect data on Android, iOS, and many other devices, online or offline, in any language and with complex skip logic.
- **ANALYZING DATA**: Inspect data moments after it was collected - and download it for advanced analysis in other software in Excel, CSV, KML, and other formats.

Sign up and first login

Visit https://kobo.unhcr.org and create a new account. After your account activation, through the link that was sent to you, you can log in to access your account.

Note that the server is open to non-UNHCR users.

Upload your form

You can now go to project and upload the form designed throught the question library in XlsForm format.

Create users to collect data

More support on KoboToolBox is avaialble form this page

Share projects

Be cautious!

Be very cautious with questionnaire versioning on a kobotoolbox project – i.e. do avoid to use a versioned project for production – you may destroy the integrity of your data set– Versioning can be used in a "production project" only to add a modality to an existing question –

Data Entry on KoboToolBox

IMPORTANT

This chapter is not written yet.

Managing Android Devices

The software platform used for the data consolidation has specific clients developped on the Android Mobile Operating System. It's also possible, although not first recommended to use any browser (including any browser on a smartphone) to record information.

A certain number of smartphones or tablets will be needed for each survey team to have one or two phones, as well as a few backups.

For ease of use, smartphones and tablets with a large touchscreen and slide-out keyboard are preferable. Equipement might be borrowed from UNHCR HQ, or country operations could have their own set of phones.

Minimal Requirements for the devices

- Android Version 2.3 (recent phones are generally platform 4 or higher)
- Screen size of 4 inches is recommended
- GPS Chipset
- Wifi connection

Required Applications:

- Latest recommended version of KoboCollect
- GPS Test, if you need GPS coordinates for the survey

Steady power supply

Access to a good electrical supply is essential for the router and computer and for over-night charging capacities of the phones. Where electrical supply is unstable with unexpected power-cuts or planned power savings, alternative charging options must be considered before the survey.

USB Battery packs

Even though they cannot replace steady electrical power, battery packs are a useful secondary source of energy for the phones. We recommend having at least 2 for a survey where the phones are daily charged. This way, if a phone lack battery during the survey, the enumerator can charge it while continuing the survey.

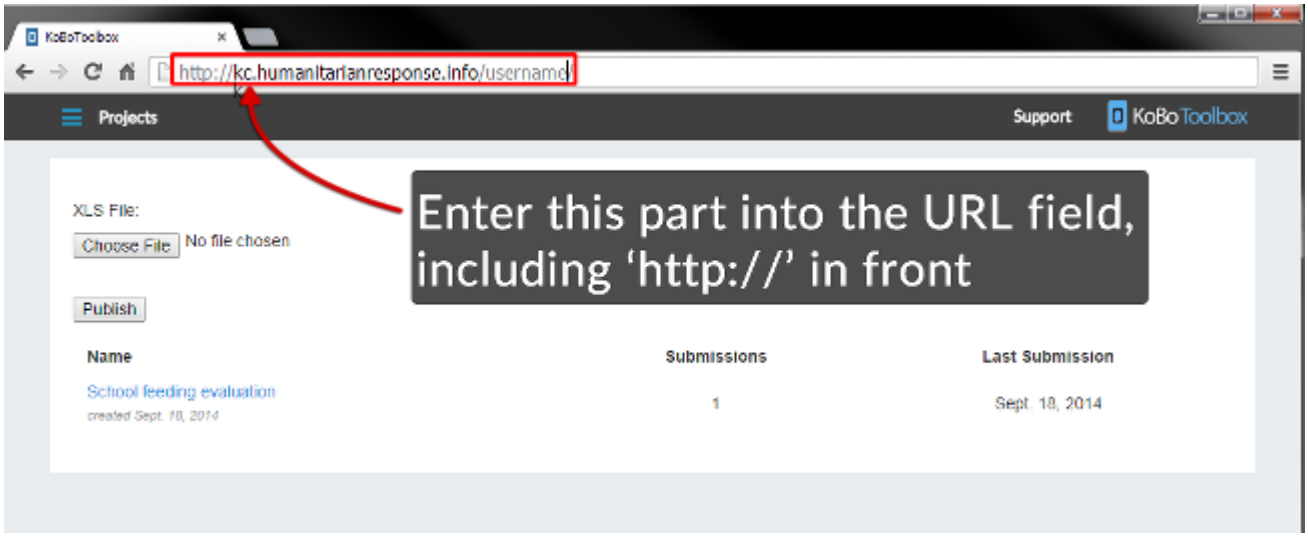
Using the client on Mobile Devices: koboCollect

KoBoCollect is an Android app that can be installed on any standard Android phone or tablet. To download the app to your Android device, click here.

Server URL set up

After installing KoBoCollect, you need to configure it so that it can be used together with your KoBoToolbox account for data collection. To do so, follow the steps:

- In KoBoToolbox, click on Projects within the menu sidebar on the left. Note the URL that is inside your browser at the top of your screen (for example, http://kobo.unhcr.org/username/)

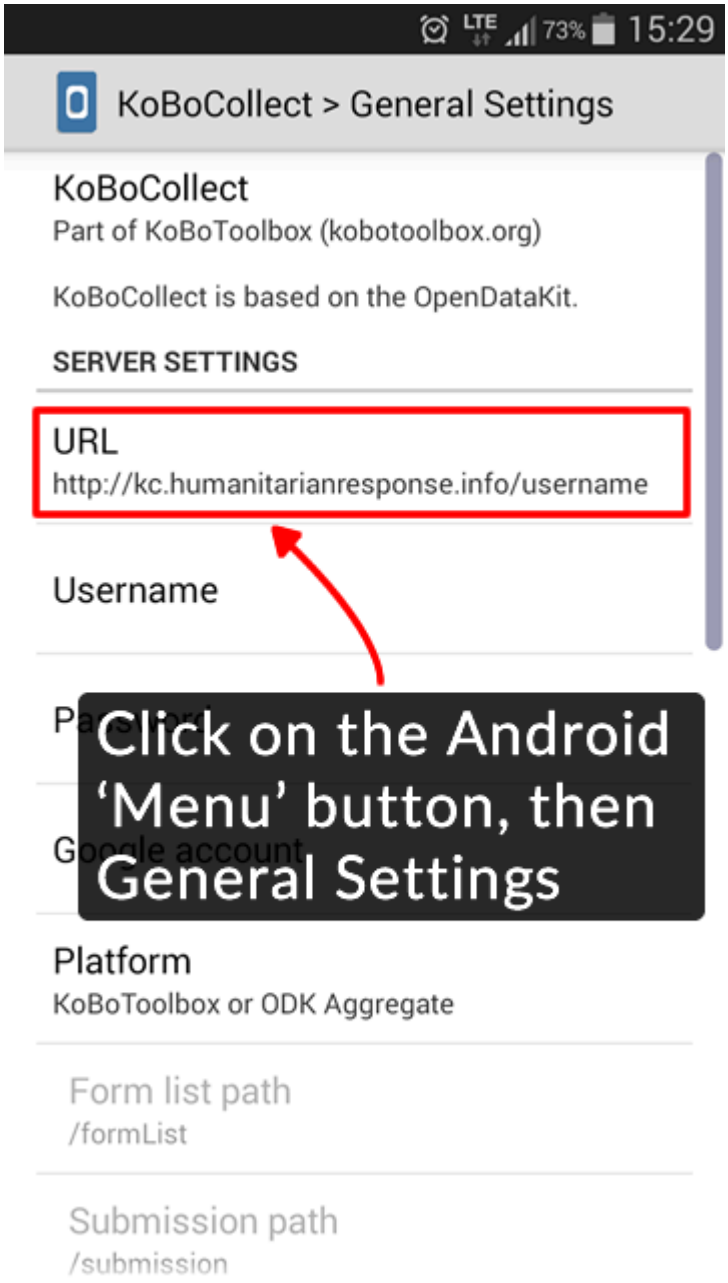


Offline

- On your Android device, open KoBoCollect and open the General Settings (click on the settings button of your device to access the settings).

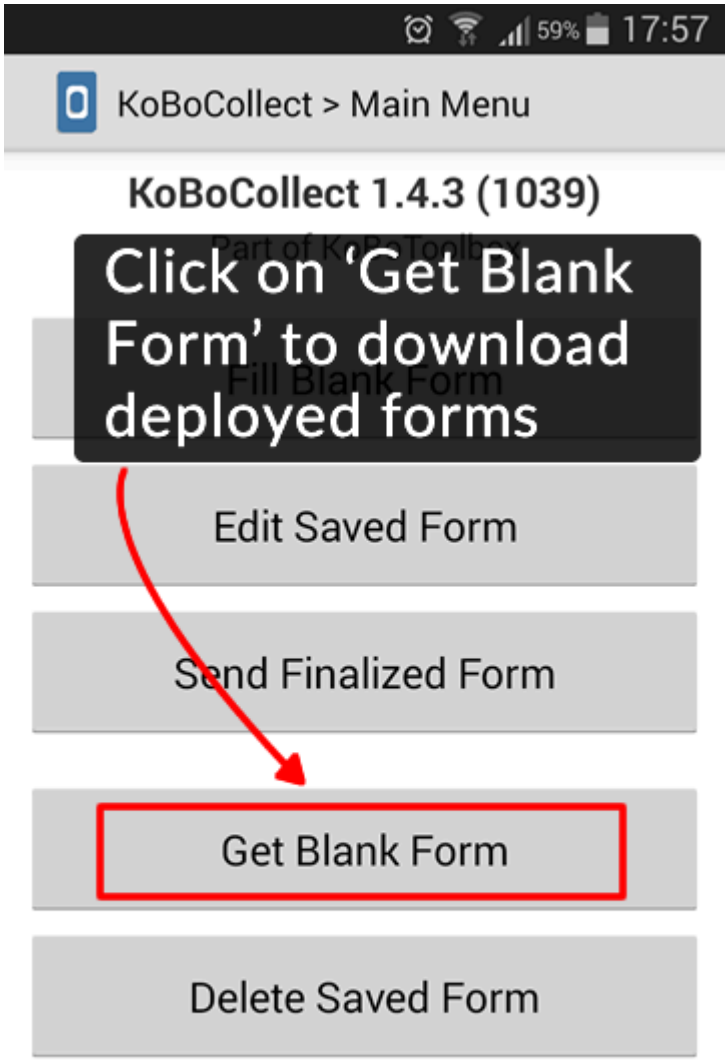
Offline

- In General Settings, under URL, enter the exact URL from step (2). Make sure you include the correct 'https://'



Offline

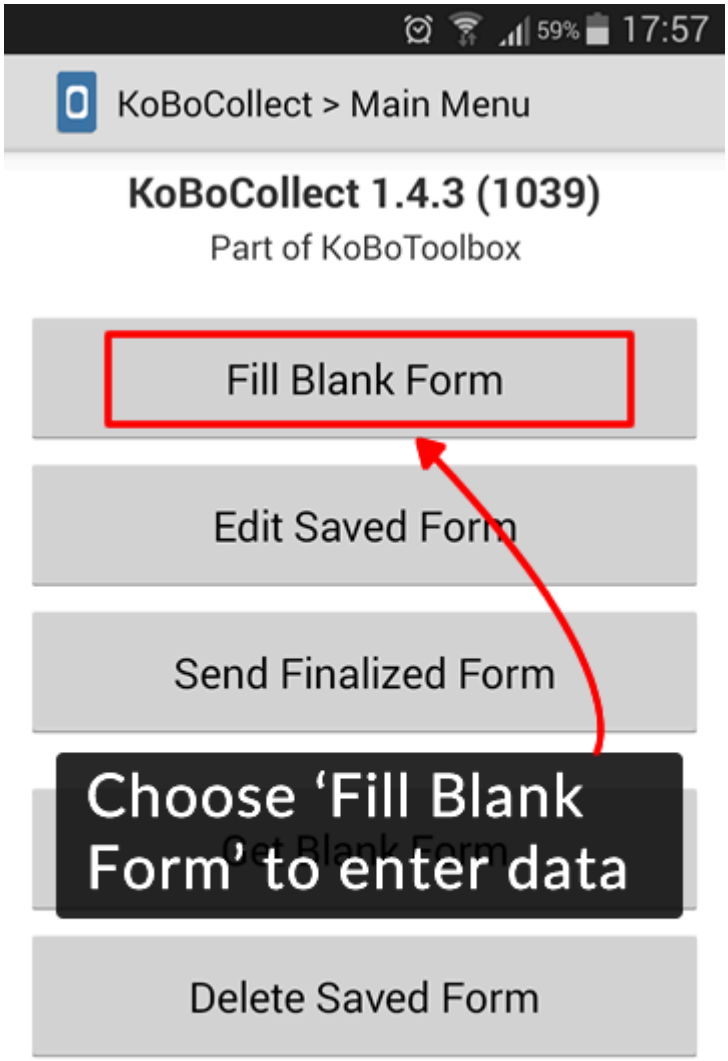
- Download forms from your account. Make sure you are connected to the Internet on your device. Also, you need to have deployed at least one project in KoBoToolbox. On the home menu of KoBoCollect, click Get Blank Form. A list of all your forms from your different projects will be shown. Click Toggle All (or select the ones you wish to download), then click Get Selected.



Offline

Collect data

- Click on Fill blank form

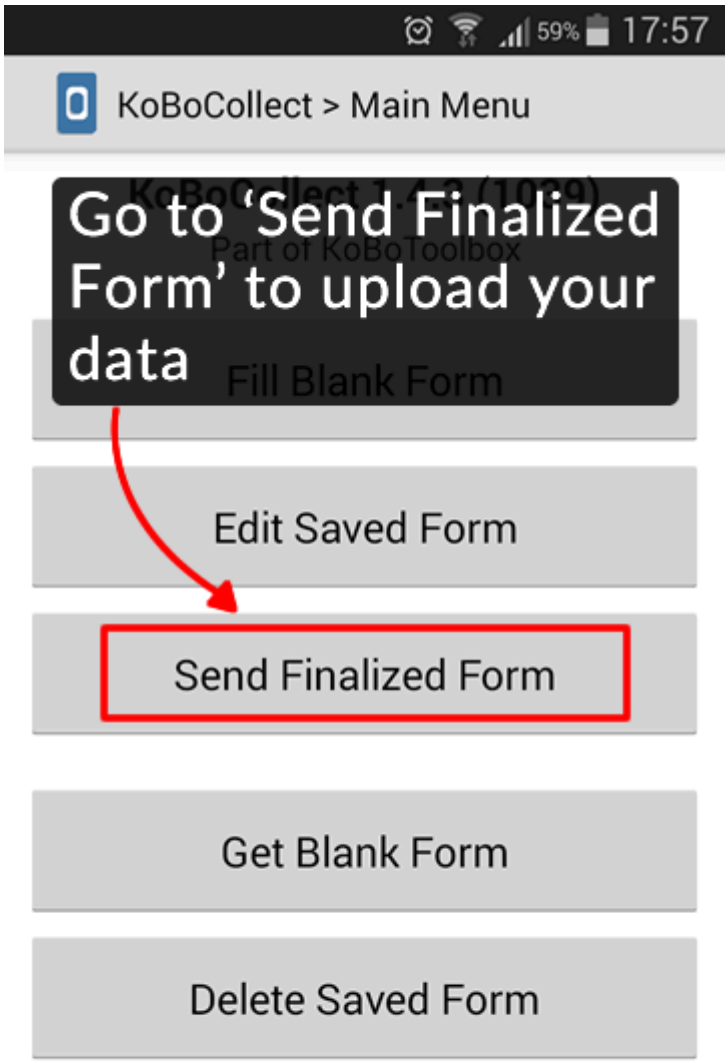


Offline

- Select the form to which you would like to enter data
- Go through all the questions (swiping your finger from right to left)
- At the end click on Save Form and Exit (making sure the form is marked as 'finalized')

Uploading finalized data

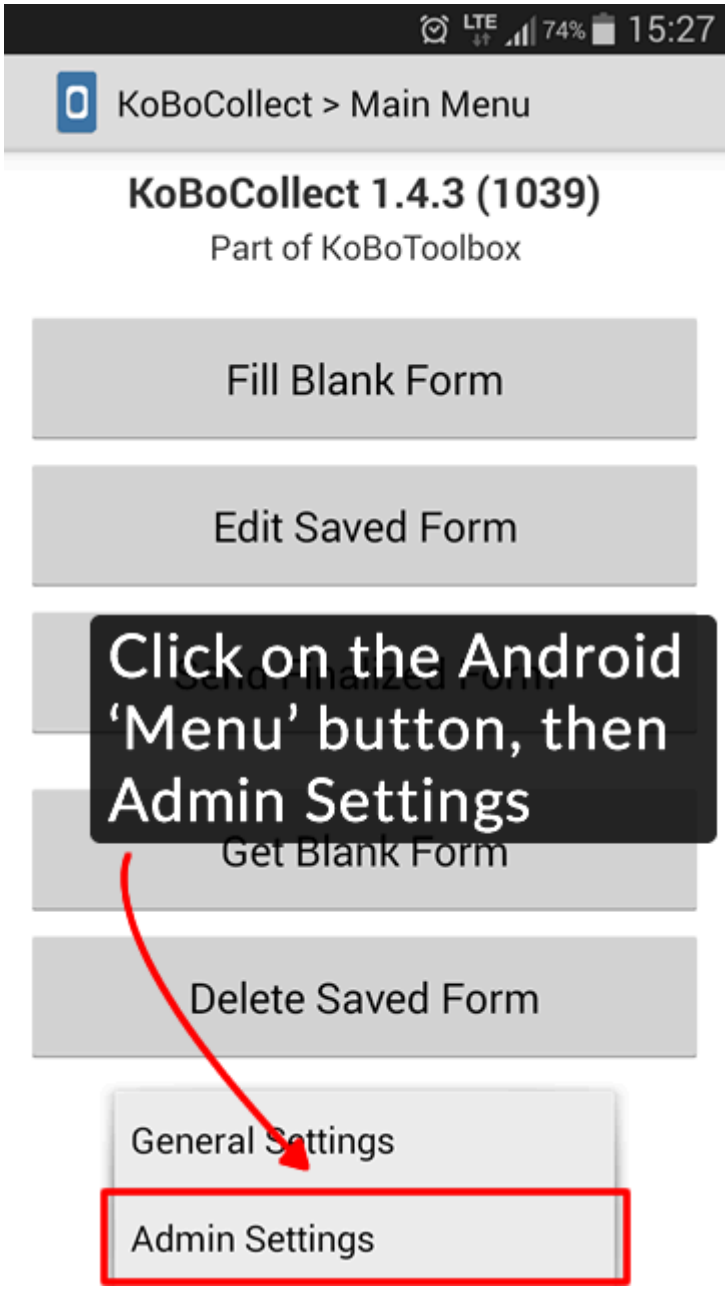
- From the home screen, click on Send Finalized Form



Offline

- A list of your most recently collected forms appears.
- Click Toggle all (or select the ones you wish to send), then click Send Selected.

It is possible to hide buttons and options within KoBoCollect. On the home screen click the Android menu button, then Choose Admin Settings and select the buttons you would like to hide from the different screens. If you set an admin password, interviewers won't be able to access the Admin settings to ever get access to these buttons.



Offline

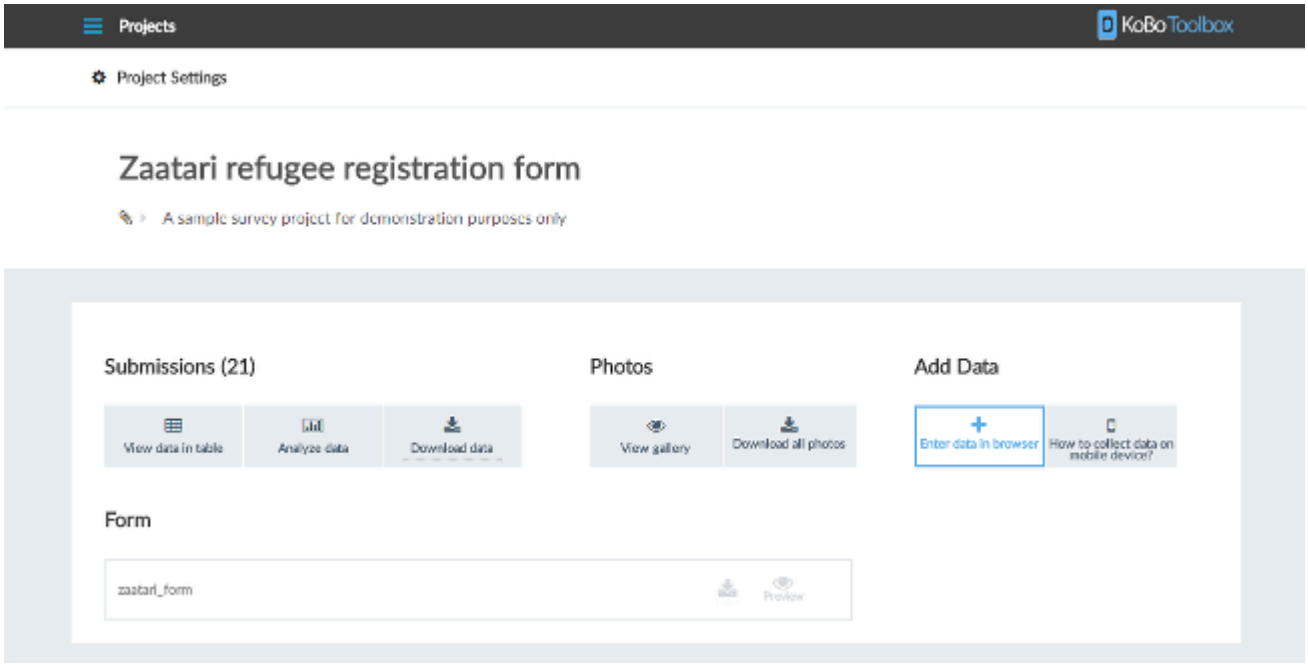
Using the web client: Enketo

Web Forms, also known as Enketo, are used by KoBoToolbox to preview forms and to enter data directly on a computer. Web forms also for collecting data on any mobile devices - even when offline at the time of data collection. It works on virtually any device, including iPhones, iPads, or any other smartphone, tablet, or computer. Some features are still being actively developed for Enketo, so some special questions may not be fully supported yet on every device. ### Start Collection

To start collecting data using web forms in your computer, simply click on the button 'Enter data in browser' in your project view. This is especially useful for testing purposes, but also when transcribing data from using paper forms.

Offline

To collect data using a mobile device, you need to copy the URL of your webform to your mobile device. You can simply send it by email or text message to any device. To obtain the URL of the webform, click either the 'Enter data in browser' or 'How to collect data on mobile device' buttons.

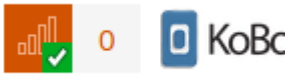


Offline

Collecting data offline

Enketo is also able to collect data while offline. However, it is essential to visit the URL once before going offline, and then saving it on the device (for example with a bookmark on the mobile browser).

Once your form has been fully loaded and cached, an offline availability icon (empty "signal bars" and a check mark) in the top-left corner will appear and indicate that the form can be accessed offline unless *browser's data are cleared*.



Offline

Enketo will display the form within that URL even without any Internet connectivity, and data will be store and queued until the next internet connection.

Differences between the two options

A detailled list of differences between the 2 options is here but in the differences are minimal and mostly impact the ammount of hidden metdata that can be collected as well the configuration of the look'n'feel of the form for end users.

Instructions for Interviewer

Role of the interviewer

The interviewer is really the "jack-of-all-trades" in survey research. The interviewer's role is complex and multifaceted.Success, therefore, depends on the quality of the interviewers' work. It includes the following tasks:

- Locating the structure and households in the sample that are assigned to them, and administering the questionnaires
- Identifying all the eligible respondents
- Interviewing all the eligible respondents in the households assigned to them
- Checking completed interviews to be sure that all questions were asked
- Making re-visits to interview respondents who could not be interviewed during the first or second visit due to various reasons
- Ensuring that the information given is correct by keeping the respondent focused to the questions
- Including their specific observations or notes on the last page of each questionnaire
- Preparing additional notes for the field editor and supervisor on other problems or observations

Locate and enlist cooperation of respondents

The interviewer has to find the respondent. In door-to-door surveys, this means being able to locate specific add resses. Often, the interviewer has to work at the least desirable times (like immediately after dinner or on weeke nds) because that's when respondents are most readily available.

Motivate respondents to do good job

If the interviewer does not take the work seriously, why would the respondent? The interviewer has to be motivated and has to be able to communicate that motivation to the respondent. Often, this means that the interviewer has to be convinced of the importance of the research.

Clarify any confusion/concerns

Interviewers have to be able to think on their feet. Respondents may raise objections or concerns that were not anticipated. The interviewer has to be able to respond candidly and informatively.

Observe quality of responses

Whether the interview is personal or over the phone, the interviewer is in the best position to judge the quality of the information that is being received. Even a verbatim transcript will not adequately convey how seriously the respondent took the task, or any gestures or body language that were evident.

Conduct a good interview

Last, and certainly not least, the interviewer has to conduct a good interview! Every interview has a life of its own. Some respondents are motivated and attentive, others are distracted or disinterested. The interviewer also has good or bad days. Assuring a consistently high-quality interview is a challenge that requires constant effort.

Approaching households

Appointment for interview

At the start of each survey cycle, either phone calls or notification letters are used to explain the purpose of the survey to the sampled households, to seek their co-operation, request the presence of all household members, collect their precise address and set up the appointment. The enumerator will then visit the households concerned in person to collect the required information. When they are not able to contact the households concerned during their visit to their quarters, a pre-printed non-contact (NC) slip will be left to these households so that another appoitnement can be set up.

Defining members of household

On order to make a comprehensive list of individuals connected to the household, the following probe approach is used:

1. First, get the name and details of the head of household, including his refugee registration number.
2. Second, get the names of all the members of your immediate family who normally live and eat their meals together here. Names, sex, and relationship to household head are first listed. For each member, in addition of getting Individual ID, the enumerator should ask if they are registered under the same UNHCR case ID than the head of household (if not get the other number and the reason why they are living together).
3. Third, get the names of any other persons related to you or other household members - "Extend Household"- who normally live and eat their meals together here. "Are there any other persons not here now who normally live and eat their meals here? for example, household members studying elsewhere or traveling". get their details and their refugee ID.
4. Then, get the names of any other persons not related to you or other household members- "Composite Household"-, but who normally live and eat their meals together here, such as servants, lodgers, or other who are not relatives. Do not list servants who have a household elsewhere, and guests who are visiting temporarily and have a household elsewhere.

Eligible Respondents for the Household Questionnaire

In each sampled household you visit, you should begin by interviewing a knowledgeable adult member of the household to fill in the Household Questionnaire. All modules of the Household Questionnaire will be administered to this person, referred to as the Household Respondent, including the modules in the questionnaire where the information collected is about other household members. The Education module is one such example.

For the purposes of the Household Questionnaire, an adult is defined as someone age 15 years and over. However, young adults (below age 18) may not be the most ideal members to interview. Therefore, in cases when there is another older household member (for instance, the parent of the 15 year-old) available to interview, you should prefer to interview this person who is likely to be more knowledgeable about the household. Whenever possible, you should use your preferences to interview the household member who is likely to be more knowledgeable.

On the other hand, interviewing the household head is not a requirement and you are not expected to ask for the household head to do the interview.

You should also keep in mind that for practical reasons, it may be an advantage to begin the Household Questionnaire with a mother or primary caretaker (of a child under five years of age), since many of the questions/modules are about children, and mothers/caretakers provide more accurate responses to such questions better than anybody else. While you should not make a special effort to ensure this, you will indeed start the interview with such persons in many cases, since, in practice, these persons are more likely to be at home than, say, male household heads.

There should only be one respondent to the Household Questionnaire and the other members of the household should not respond to any part of the questionnaire. Multiple respondents to the questionnaire will undoubtedly lead to an uncontrolled, low quality interview, and may lead to errors in recording responses. Ideally, the household respondent is not expected to consult other members that may be available in the household. However, you may allow the household respondent to ask other members in order to get more correct information, especially on information such as age, which may affect the eligibility of some members for individual questionnaires, or modules where age checks are important, such as the education, child labour, or the child discipline modules.

Eligible Respondents for the Individual Questionnaires

When you have completed the Household Questionnaire, you will have identified women (age 15-49), men (age 15-49) and 'mothers or primary caretakers' (age 15 or above) of children under five to whom you

or other interviewers in your team will administer the individual questionnaires.

- You should interview separately all women age 15 through 49 who reside in the household to fill in the Questionnaire for Individual Women.
- You should interview separately all men age 15 through 49 who reside in the household to fill in the Questionnaire for Individual Men.
- You should administer the Questionnaire for Children Under Five to mothers of children under 5 years of age who are residing in the household. If the mother is not recorded in the List of Household Members (if the mother is not a member of this household), then the person who is acknowledged by the household respondent as the primary caretaker should be the respondent to the Questionnaire for Children Under Five.

You will identify these individuals by completing the List of Household Members in the Household Questionnaire.

If you visit a household where there are no members eligible for the individual questionnaires, you must still ask questions about the household to a knowledgeable adult household member and complete the Household Questionnaire.

As a general rule, the respondent to any of the questionnaires must be at least 15 years old. This also applies to the mother or primary caretaker of a child under age 5; in the rare event that a mother or primary caretaker is less than age 15 you should record 'Other' as result of the interview in UF9 and specify that the mother/caretaker is less than age 15 and therefore cannot be interviewed. No other respondent is permitted than the mother/caretaker identified in the List of Household members.

Finding and Re-Visiting Households

Your supervisor will give you a list or tell you how to find the households to visit. You must visit all these households and should not replace these households with other households that are not selected for interviews.

If no one is at home when you go to interview the household, ask the neighbours whether anyone lives at this location. If it is occupied, ask the neighbours when the household members will return. Arrange with your supervisor to go back to the location when the household members are expected to be back; for example, at the end of the day. Note such plans on your cluster control sheet and note the time you are to return on the first page of the questionnaire (Household Information Panel).

If no adult household member is at home, arrange to come back at another time. Do not interview a household member younger than age 15, a temporary caretaker of the children, such as a daytime babysitter, and do not interview anyone who does not usually live in the household. The rule to interview a knowledgeable adult household member cannot be relaxed or violated under any circumstances.

Each household in the sample has to be visited at least three times (two re-visits) before you can mark HH9 (Result of household interview) as 'No household member or no competent respondent at home at time of visit', unless otherwise instructed by your supervisor. There may be cases when you learn that

the household will be away for an extended period, and will definitely not return within the fieldwork period, in which case HH9 would be marked as 'Entire household absent for extended period of time'. In such cases, three visits to the household may not be necessary. However, even in such cases, the ultimate decision will have to be taken by your supervisor.

If an eligible woman or man, or a mother or primary caretaker is not available for the individual interview or is not at home, ask a household member or neighbour to find out when she/he will return. Note this on the Woman's, Man's or Under-5's Information Panel, follow your supervisor's instructions, and return to interview her/him at that time. Do not take responses for these questionnaires from anyone other than the eligible person her/himself.

The person to be interviewed for the Questionnaire for Children Under Five should be the mother. A person other than the mother of the child under five can be interviewed only if the mother is living elsewhere or is deceased, and therefore does not appear in the List of Household Members in the Household Questionnaire. In these cases, the person who is acknowledged by the household respondent as the primary caretaker of the child in that household should be interviewed. If the mother/primary caretaker is not available for interview or not at home, try to find out when she/he will be available and return to the household later. If the person will not be available or will not return home at a time later that day when it is feasible to interview her/him, follow the instructions of your supervisor about the number of times you should attempt the interview.

If a child under five is not available, but the mother/primary caretaker is available, you can complete the Questionnaire for Children Under Five, with the exception of the Anthropometry module, since you need the child to perform measurements. In such a case, complete the questionnaire with the mother/primary caretaker, but leave the Anthropometry module blank to be completed during the next visit. Note this and discuss with your supervisor. If the child is still not available after the re-visit(s), record the result in question AN2 as 'Child not present'. Re-visits should be planned by supervisors, if possible, to measure the heights and weights of children, when children are not present at the time of first visit to the household.

Ask your supervisor if you are in doubt about what to do when you cannot locate a household, or you cannot complete an interview. Always keep a record on the cluster control sheet of the households you visited where nobody was at home. If it is not possible to interview an eligible woman or man, record this on the Woman's or Man's Information Panel of the respective questionnaires. If it is not possible to interview a mother or primary caretaker, record this on the Under Five Child Information Panel of the Questionnaire for Children Under Five.

How to handle the interview

The interviewer and the respondent are strangers to each other; therefore, one of the main tasks of the interviewer is to establish rapport with the respondent. The respondent's first impression of you will influence her/his willingness to participate in the survey. Make sure that your appearance is neat and you also appear friendly as you introduce yourself.

On meeting the respondent, the first thing you should do is to introduce yourself, stating your name, the

organization you are working for, the objectives of the survey, and what you want the respondent to do for you. You are advised to avoid long discussions on issues which are not related to the survey and which may consume a lot of your time.

After building rapport with the respondent, ask questions slowly and clearly to ensure the respondent understands what he/she is being asked. After you have asked a question, pause and give the respondent time to think. If the respondent feels hurried or is not allowed to form his/her opinion, he/she may respond with "I don't know" or give an inaccurate answer.

Specifically, the following guidelines will help you handle interviews:

- Ensure that you understand the exact purpose of the survey and each question. This will help you to know if the responses you are receiving are adequate or relevant.
- Remember the survey schedule, and remember that you are part of a team. Do not stay and talk for too long, but do not rush the interview either.
- Ask the questions exactly as they are written. Even small changes in wording can alter the meaning of a question.
- Ask the questions in the same order as they are given on the questionnaires. Do not change the sequence of the questions.
- Ask all the questions, even if the respondent answers two questions at once. You can explain that you must ask each question individually, or say "Just so that I am sure…" or "Just to refresh my memory…", and then ask the question.
- Help your respondents feel comfortable, but make sure you do not suggest answers to your questions. For example, do not 'help' a woman remember various contraceptive methods. Those cases when you are expected to 'help' the respondent, such as probing for answers or using information to remind the respondent of dates, ages, and durations are clearly indicated on the questionnaires, and are topics that are covered during your training.
- Do not leave a question unanswered unless you have been instructed to skip it. Questions left blank are difficult to deal with later. When questionnaires arrive at the central office for editing and data entry, it may look as though you forgot to ask the question. Always write in '0' when a zero answer is given. For some questions, the code 'DK' will already be provided, and after you are sure that the respondent is unable to provide you with an answer, you will be able to circle this response. In questions where a 'DK' response is not printed on the questionnaire, you must make sure that the respondent provides an answer. In exceptional cases where this may not be possible, indicate this on the questionnaire with a note.
- Record answers immediately when the respondent gives you the responses. Never rely on writing answers in a notebook for transfer to the questionnaire later.
- Check the whole questionnaire before you leave the household to be sure it is completed correctly.
- Thank the respondent for her/his cooperation and giving you time to interview her/him. Leave the way open to future interviews (for re-visits). Avoid over-staying in the respondent's household even if he/she is very friendly and welcoming.

General Points

- **Make a good first impression**

The first impression a respondent has of you is formed through your appearance. The way you dress may affect whether your interview is successful or not. Dress neatly and simply.

When first approaching the respondent, do your best to make her/him feel at ease. With a few well-chosen words, you can put the respondent in the right frame of mind for the interview. Open the interview with a smile and greetings and then proceed with your introduction as specified on your questionnaire.

If and when necessary, tell the respondent that the survey will help the development of plans for children and women and that his/her cooperation will be highly appreciated.

- **Gain rapport with the respondent**

Try not to arrive at the selected household at an inconvenient time of day, such as mealtimes, or too late or early during the day. Try to arrive when the respondents will not be too busy to answer questions.

Introduce yourself by name and show your identification. Explain the survey and why you want to do interview in the household, exactly as your introduction tells you to.

Be prepared to explain what is meant by confidentiality and to convince respondents to participate if they are reluctant.

Make sure that the respondents do not confuse you with others who might be visiting households for other reasons; for instance, for selling goods.

If the respondent refuses to be interviewed, note the reasons on the questionnaire, if possible.

- **Remain calm and polite at all times.**
- **Always have a positive approach** Never adopt an apologetic manner, and never approach with such words as “Are you too busy?”. Such questions will invite refusal before you start. Rather, tell the respondent “I would like to ask you a few questions”.
- **Stress confidentiality of information collected** Always stress confidentiality of the information you obtain from the respondent. Explain to the respondent that the information you collect will remain strictly confidential and that no individual names will be used for any purposes, and that all information will be grouped together and depersonalized when writing the report. Use a language understandable by the respondent to get this message across. Never mention other interviews or read the questionnaire with other interviewers, the editor or the supervisor in front of a respondent or any other person. This will automatically erode the confidence the respondent has in you.
- **Probe for adequate responses** You should phrase the question as it is in the questionnaire. If you realize that an answer is not consistent with other responses, then you should seek clarification

through asking indirect questions or some additional questions so as to obtain a complete answer to the original question. This process is called probing. Questions, while probing, should be worded so that they are neutral and do not lead the respondent to answer in a particular direction. Ensure the meaning of the original question is not changed.

Pause and wait if the respondent is trying to remember difficult items.

Ask the respondent to clarify her/his answer if necessary. You may have misunderstood the response.

Check for consistency between the answers a respondent gives. Treat the questionnaires as tools that you are using to converse with the respondent. Try to understand and remember the responses, and if there is an inconsistency, ask the questions again. However, never point out to the respondents inconsistencies that you may have identified in a manner that may be understood as if you are testing the respondent's honesty or integrity.

- **Answering questions from respondent** The respondent may ask you some questions about the survey or how he/she was selected to be interviewed or how the survey is going to help her/him, before agreeing to be interviewed. Be direct and pleasant when you answer. The respondent may also be concerned about the length of the interview. Please be frank to tell him/her how long you are likely to take to administer the questionnaire.

- **Interview the respondent alone** The presence of a third person during the interview can prevent you from getting frank and honest answers from the respondent. It is, therefore, very important that the interviews are conducted privately and that all the questions are answered by the respondent only. This is especially important in the case of the Woman's and Men's Questionnaires, which include several topics that the respondents will consider to be "personal" or "private". If other people are present, explain to the respondent that some of the questions are private and request to talk to her/him while alone.

- **Handling hesitant respondents** There may be situations where the respondent simply says "I don't know", or gives an irrelevant answer or acts in a manner suggesting he/she is bored or contradicts earlier answers. In all these cases, try your best to make him/her get interested in the question. Spending a few moments to talk about things unrelated to the interview (e.g. his/her town or village, the weather, his/her daily activities etc.) may be useful.

- **Adopt a non-judgemental attitude** "Social desirability response bias" is a potential problem in surveys and refers to the tendency for respondents to present a favourable image of themselves to the interviewers. Sensitive questions may lead respondents to adjust their answers so as to appear politically correct or socially acceptable. Questionnaire items with strong social norms (such as adherence to religious or cultural expectations), or adopting attitudes/activities/objects that are widely considered desirable or undesirable tend to elicit "socially acceptable answers" rather than correct and honest answers. To minimise social desirability response bias it is very important to adopt a non-judgemental attitude and to not display any of your own attitudes, such as cultural or religious values, political preferences, and the like.

Instructions for Supervisors & Editors

IMPORTANT

This chapter is not written yet.

Responsibilities

The field supervisor is the senior member of the field team. He/she is responsible for the well-being and safety of team members, as well as the completion of the assigned work and the maintenance of data quality.

The specific responsibilities of the field supervisor are to make the necessary preparations for the fieldwork, to organize and direct the data collection in his/her assigned clusters, and to spot check the data collected in especially the Household Questionnaire.

Preparing for fieldwork requires that the field supervisor: (1) Obtains sample household lists and maps for each area in which his/her team will be working, discuss any special issues, such as potential security conditions in certain areas. (2) Becomes familiar with the area where the team will be working and determine the best arrangements for travel and accommodations. (3) Contacts local authorities to inform them about the survey and to gain their support and cooperation. (4) Obtains all monetary advances, supplies and equipment necessary for the team to complete its assigned interviews.

Task

Careful preparation by the field supervisor is important for facilitating the work of the team in the field, for maintaining interviewer morale and for ensuring contact with the central office throughout the fieldwork.

Organizing fieldwork requires that the field supervisor: (1) Assigns work to interviewers, taking into account the linguistic competence of individual interviewers, and assures that there is an equitable distribution of the workload. (2) Coordinate the work of the measurer by making sure he/she knows where to find the households that interviewers are conducting interviews in and approximately how many children and at what time a visit to the household should happen. (3) Maintains Cluster Control Sheets, and makes sure that assignments are carried out. (4) Makes spot checks of the Household Questionnaire (and individual questionnaires when appropriate) by conducting interviews according to the procedure described below. (5) Regularly sends completed questionnaires and progress reports to the fieldwork director and keeps headquarters informed of the team's location. (6) Communicates any problems to the fieldwork director. (7) Takes charge of the team vehicle, ensuring that it is kept in good repair and that it is used only for project work. (8) Makes an effort to develop a positive team spirit. A congenial work atmosphere, along with careful planning of field activities, contributes to the overall quality of a survey.

Monetary Advances for Field Expenses

The field supervisor should have sufficient funds to cover expenses for the team. Funds for team members should be distributed according to the procedures established by the survey director, if these have not been included in the per diem that is given directly to the interviewers.

The field supervisor should arrange for a system to maintain regular contact with the central office staff before leaving for the field. Regular contact is needed for supervision of the team by central office staff, payment of team members, and the return of completed questionnaires for timely data processing.

Arranging transportations & accomodations

It is the field supervisor's responsibility to make all necessary travel arrangements for his/her team, whenever possible, in consultation with the central office. The field supervisor is responsible for the maintenance and security of the team vehicle. The vehicle should be used exclusively for survey-related travel, and when not in use, should be parked in a safe place. The driver of the vehicle takes instructions from the field supervisor.

> VEHICLES ARE GENERALLY PROVIDED TO TRANSPORT THE TEAM TO ASSIGNED WORK AREAS. HOWEVER, IN SOME CASES, IT MAY BE NECESSARY TO ARRANGE FOR OTHER MEANS OF TRANSPORTATION, SUCH AS BOATS, HORSES, MULES, ETC. CUSTOMIZE THE PARAGRAPH ABOVE ACCORDINGLY.

In addition to arranging transportation, the field supervisor is in charge of arranging for food and lodging for the team. If they wish, interviewers may make their own arrangements, as long as these do not interfere with fieldwork activities. The lodging should be reasonably comfortable, located as close as possible to the interview area, and should provide a secure space to store survey materials. Since travel to rural clusters is often long and difficult, the field supervisor may have to arrange for the team to stay in a central location.

Contacting local authorities

It is the field supervisor's responsibility to contact the regional, district, local, and village officials before starting work in an area. Letters of introduction will be provided, but tact and sensitivity in explaining the purpose of the survey will help win the cooperation needed to carry out the interviews.

Using maps to locate clusters

> THIS SECTION PRESUMES THAT A FRESH HOUSEHOLD LISTING HAS BEEN CONDUCTED AND THAT UPDATED CLUSTER LOCATION AND SKETCH MAPS THEREFORE ARE AVAILABLE TO TEAMS.

A major responsibility of the field supervisor and the field editor is to assist interviewers in locating households in the sample. The fieldwork director will provide the supervisor with a copy of the Household Listing for the sample as well as base, location, and sketch maps of the clusters in which his/her team will be working. These documents will enable the team to identify the cluster boundaries and to locate the households selected for the sample. The representativeness of the sample depends on finding and visiting every sampled household.

Maps are generally needed during all stages of a survey, since they provide a picture of the areas in which interviews are to be carried out and help to eliminate errors, such as duplication or omission of areas. Moreover, maps help the team determine the location of sample areas, the distance to them, and how to reach selected households or dwellings.

Each team will be given general base and location maps, Household Listing Forms, and sketch maps, and written descriptions of the boundaries of selected areas. A cluster (i.e., PSU or EA) is the smallest working unit in any census or survey operation that can easily be covered by one enumerator. It has identifiable boundaries and lies wholly within an administrative or statistical area. The general base maps will show more than one cluster. Each cluster is identified by a number (for example, EA-010400105). Symbols are used to indicate certain features on the map such as roads, footpaths, rivers, localities, boundaries, etc. If symbols are shown on the map, the field supervisor and field editor should know how to interpret them by using the legend.

In most clusters, the boundaries follow easily recognizable land features such as rivers, roads, railroads, swamps, etc. However, at times, boundaries are invisible lines. The location and determination of invisible boundaries calls for some ingenuity, particularly in rural areas. If the location and sketch maps and descriptions do not provide enough detail, the following procedure is suggested:

In rural areas:

1. Identify on the map the road used to reach the cluster. When you reach what appears to be the cluster boundary, verify this by checking the location of actual terrain features and landmarks against their location on the map. Do not depend on one single feature; rather, use as many as possible.
2. It is usually possible to locate unnamed roads or imaginary lines by asking people living in the vicinity. In most cases, these people will know where the villages are and, by locating the villages, you can usually determine where the boundaries run. Local authorities may be helpful, as well as residents.
3. While there are cases in which boundaries shown on the map no longer exist (for example, they have been demolished), or have changed location (for example, a road has been relocated or a river has changed course), do not be hasty in jumping to conclusions. If you cannot locate a cluster, go on to the next one and discuss the matter later with the fieldwork director.

In urban areas:

1. There should be no problem with invisible lines, as urban areas generally have plenty of boundaries for use.
2. Street names in urban areas will often help you to locate the general area of clusters. Boundaries can be streets, alleys, streams, city limits, power cables, walls, rows of trees, etc.
3. Check the general shape of the cluster. This will help you find out if you are in the right place.
4. Read the written description.
5. You should locate all the cluster boundaries before you begin interviewing. For example, if the cluster is a rectangular block, the names of three boundary streets is not enough to unequivocally identify the cluster; check all four boundary streets.

Instructions for Managers

Observation and supervision throughout the fieldwork are a part of the training

Team supervisors play very important roles in continuing this training and in ensuring the quality of data

In addition, supervisors are responsible for the organisation of daily work and for the security of staff and equipment.

Notes from the Field: How to incentivize your survey team

Spot-check household composition

- Supervisors should complete parts of form and compare with that of the interviewer
- Check about 5% of households (5-6 per week)
- All team members must be spot-checked; provide feedback if necessary

OBSERVING INTERVIEWS

- To evaluate and improve interviewer performance
- To look for errors and misconceptions that cannot be detected through editing

Evaluating Interviewers Performance

- Re-read relevant sections from the Interviewer's Manual with the team to resolve problems
- Encourage the interviewers to talk about any situations they encountered in the field
- Discuss whether situations are handled properly, and how to do it in the future

What is to be monitored on daily basis

Daily montoring can be done by downloading data from kobo server and performing a few check…

- Overall completion rate of the survey according to planned logistics
- Monitor usage of __or_other_ for questions where selection of modalities is open
- Monitor usageRespondent refuses to be interviewed
- Distribution of completion by clusters when the sampling is based on cluster or on strata if stratfied sample
- Rate of work by teams
- Synchronization timing - verify Synchronize survey data

Field Check Tables

- Provide a full range of information about the quality of the data already collected as per above
- Provide information on the work of each team and each interviewer
- To be shared on a regular basis
- Be transparent and report on problems ..before others detect them

Clean & Anonymise

IMPORTANT

Data Cleaning is an essential step in data analysis. Most of the cleaning should allow to predect some potential errors in the dataset.

A multiplicity of errors can occurs during a survey. Once data are collected only a few of them can be addressed.

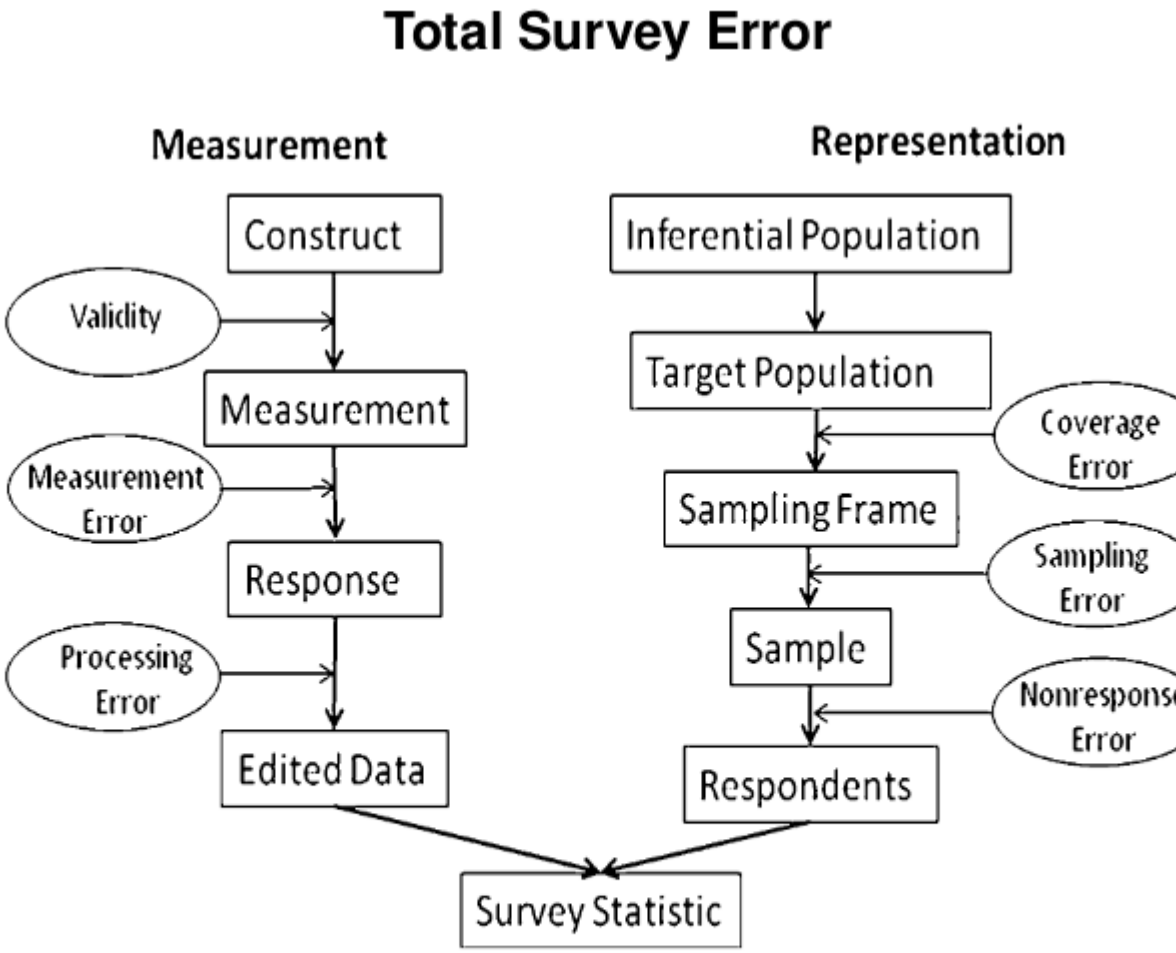


Figure 3. Total Survey Error Components Linked to Ste Measurement and Representational Inference Process (Groves e

Clean

Cleaning linked to the sample

In this case, some household were visited multiples times. The only way to address this is to have a cleaning log that contains the uniqueID (meta_instance_id) of the questionnaires that will need to be deleted.

Cleaning on numeric variable: Filter Unwanted Outliers

Outliers on numeric variable can be spotted looking at the value of standard deviation.

Keep in mind that “outliers are innocent until proven guilty”. Sometimes they traduce a proable in data collection but sometimes it can also add important information. One must have a good reason for removing an outlier, such as suspicious measurements that are unlikely to be real data.

Cleaning categoric variables: review categories

xlsform allow to add the option of using **or_other** when raising a categoric questions. Often the results of this sub-variable needs to be reviewed to ensure that the or_other can not be affected to any existing value.

The other point to check is low-frequency modalities that might need to be compacted together. This is also important in the sense that it will facilitate the data anonymisation in case this variable is a key one.

Fix labeling

When recoding the data set it can appears often that the label attributed to the variables when raising the questions are not worded well, comme with typos / capitalisation issues or are too long to appear in a report. This can be fixed directly within the xlsform used for the data analysis plan.

Handle Missing Data

This can be important when developing the analysis for the scoring.

The 2 most commonly recommended ways of dealing with missing data are:

- Dropping observations that have missing values: Dropping missing values is sub-optimal because when you drop observations, you drop information. The fact that the value was missing may be informative in itself.
- Imputing the missing values based on other observations

The best way to handle missing data for categorical features is to simply label them as ’Missing’!

- You’re essentially adding a new class for the feature.
- This tells the algorithm that the value was missing.
- This also gets around the technical requirement for no missing values.

For missing numeric data, you should flag and fill the values.

- Flag the observation with an indicator variable of missingness.
- Then, fill the original missing value with 0 just to meet the technical requirement of no missing values.

By using this technique of flagging and filling, you are essentially allowing the algorithm to estimate the optimal constant for missingness, instead of just filling it in with the mean.

Anonymisation

Describe

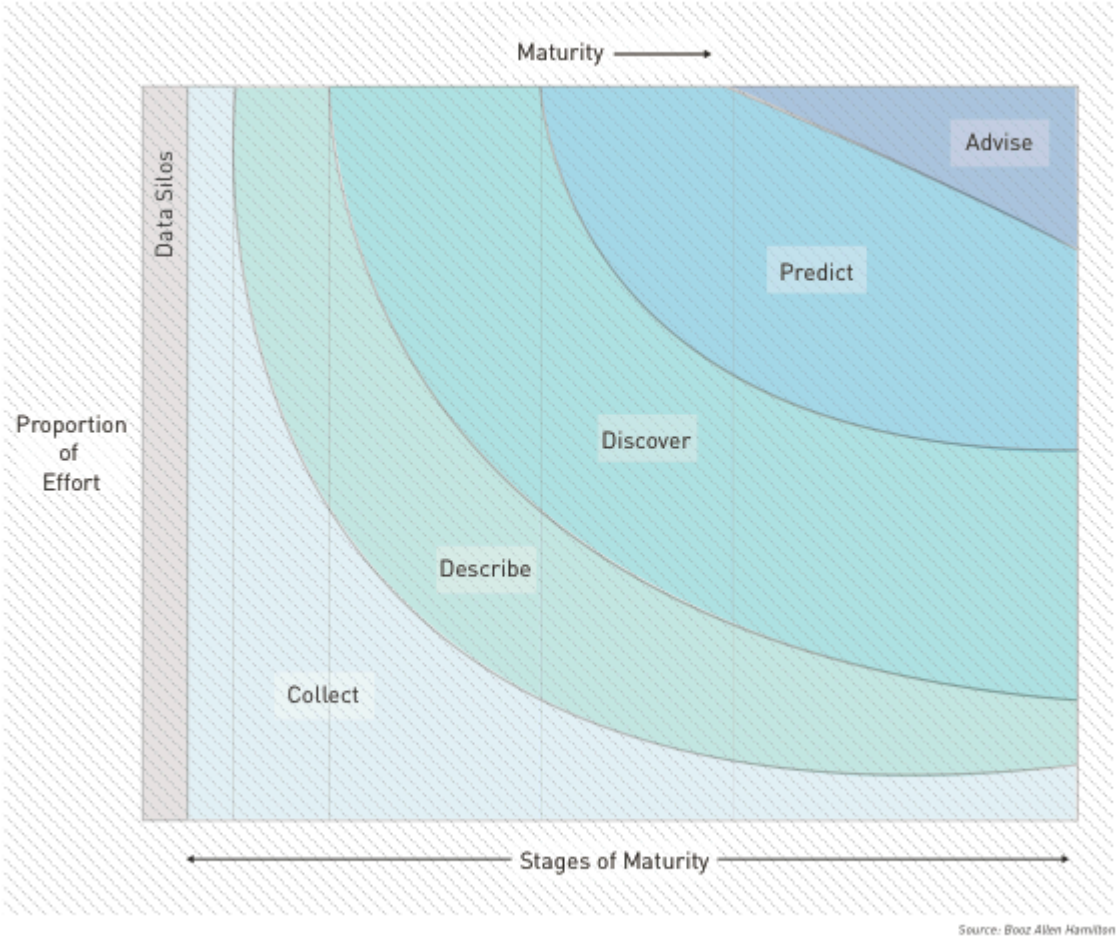
IMPORTANT

Analytics approaches usually encompasses 4 distinct steps:

- Describe;
- Discover;
- Predict;
- Advise.

The first step is extremely important as it provides a first overview of the data and allow to build the foundation for the next ones.

The data science maturity model from Booz Allen provide a simplified overview of analytics challenges.



The Data Science Maturity Model

The first phase (**DESCRIBE**) of analytics consists in a thourough review of all variables within the dataset.

This exploratory phase usually overlaps with data cleaning; it is the stage where anomalies become evident e.g. individually plausible values may lead to a way-out point when combined with other variables on a scatterplot. In an ideal situation, this step would end with confidence that one has a clean dataset, so that a single version of the main datafiles can be finalised and 'locked' and all published analyses derived from a single consistent form of 'the data'. In practice later stages of analysis often produce additional queries about data values.

Tabulation: One-way table & Crosstabulation

Often, much of the basic information need are supplied through a tabulation of results, question by question, as 'one-way tables'. Sometimes this can be done using an original questionnaire and writing on it the frequency or number of people who 'ticked each box'. Of course this does not identify which

respondents produced particular combinations of responses, but this is often a first step where a quick and/or simple summary is required.

At the most basic level, cross-tabulations break down the sample into two-way tables showing the response categories of one question as row headings, those of another question as column headings. If for example each question has five possible answers the table breaks the total sample down into 25 subgroups.

When dealing with *select_multiple*, there are different ways of computerising these data. The “multiple dichotomy”approach provides as many columns as there are alternatives. The “multiple response” way find the maximum number of ticks from anyone and then have this number of columns, entering the codes for ticked responses, one per column.

Graphical review

The use of graphical methods is important for presentational purposes, where simple messages need to be given in easily understood, and attention-grabbing form.Their true power comes from the ability of the eye to discern patterns in a graph that are not clearly evident from lists of numbers or tabulated statistics. In Tufte’s pithy phrase, “**graphics reveal data**”.

With data in hand, the most productive first step is often to explore the data graphically. These graphs do not have to be especially polished and beautiful; rather, they need to be easy to produce and thoroughly informative, a visual scratch pad where we use the power of graphics to get a sense of the shape of variables and the interactions among them.

Type of variables.

When designing the form, each variable is associated with a data type. The recommended graphical view is actually linked to the subtype of variables that should be presented. A correct initial mapping of variable will partly allow for automation of the step (cf. next article).

- **Categoric**

Categoric Subtype	**Xlsform type**	**Examples**
Nominal	select_one, select_multiple	occupation
Ordinal	select_one	education
String	text	comments

- **Numeric**

Numeric Subtype	Xlsform type	Examples
Continuous	integer, numeric, calculate	age, height
Discrete	integer, select_one, calculate	cars, floors
Date/Time	date, time, dateTime, calculate	DoB, Arrival Date

- **Geographic**

Geographic Subtype	Xlsform type	Examples
Point	geopoint, select_one	Cities
Line	geotrace, select_one	Road
Polygone	geoshape, select_one	Gov., District

Barchart

Ordered bar

Barchart shows simple frequency distribution of **categoric value**.

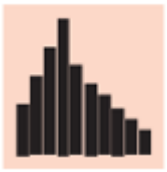
Usually, horizontal bar chart are used so that the label of modalities for the variable are legible. If the variable is nominal, bars will be ordered as per the frequency, if the variable is ordinal, will be order as defined within the variable definition.

Bar chart can also be used to visualise crosstabulation (i.e. the frequency relation between **two categoric variables**), bar chart can be be: * faceted
* stacked
* paired

Paired column

Histogramm

Histogram



Histogram shows simple frequency distribution of a **unique numeric variable**.

The classification of data into bins serves to filter out some of the noise. The key choice that has to be made when designing an histogram is that of the number of classes ("bins") into which to group the data or, alternatively, the width of each class, and this choice is as much a matter of art as of science. The Freedman and Diaconis rule or the Wand "zero-stage rule" can be used to determine bin width.

Note that histogramm can also be conveninent used to display date variables. The same discreatisation approach can be used.

Line Chart

Line

Line chart are used to display smoothen frequency distribution of a **unique numeric variable**.

A histogram provides a discretized, nonparametric approximation to the underlying density, but it has three drawbacks: it is not smooth, it depends on the bin widths, and it is sensitive to the choice of end points of the bins. in order to address those points, it's possible to use Kernel density visualisation in order to see a smoothed version of the data.

Boxplot

Boxplot

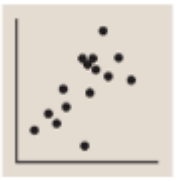
Boxplot are used to visualise the interaction between **one categoric variable and one**

numeric variable.

At the heart of the boxplot is a box that shows the 25th percentile (the "lower hinge") and 75th percentile (the "upper hinge"), as well as the median value of the observations. On the upper side of the box one adds a line that stretches to the "upper adjacent value" and ends in a whisker; the whisker extends up to the largest data point that lies within 1.5 interquartile ranges of the 75th percentile. A similar line stretches to the lower adjacent value. All values of the variable beyond the adjacent values (the "outside values") are plotted. The result is a snapshot of the distribution of a variable that allows one to get a sense of its symmetry, and the role of outliers. Boxplots are exploratory devices, and one must resist the temptation to try to use them for statistical inference.

Scatterplot

Scatterplot



Scatterplot are used to visualise the crosstabulation of **two numeric variable**.

The shape of the resulting cloud of points will allow to detect quickly potential correlation. A regression line can be automatically drawn on the scatterplot to help visualising correlations.

Scatterplots can also help detecting specific patterns that would not be detectable if just using means, variance & correlation (c.f. Anscombe'squartet).

Correlation plot

XY heatmap

Correlation plot are used to display the level of **correlation between different modalities of 2 categoric variables**.

They are often used in conjonction with statistical test (such as chi-squarred cf below)

Maps

It is often useful to display statistical information on maps and indeed a statistical map is just subtype of chart for geographic variable.

Proportional symbol

Proportion symbol map are used for absolute values (i.e. to be used to visualise crosstabulation between **a geographic variable and a numeric variable**)

Basic choropleth

Choropleth map are used to display proportion of ration (i.e. to be used to visualise crosstabulation between **a geographic variable and the frequency of a categoric variable**).

Flow map

Flow maps are used to present the **relation between two geographic variables**.

WordCloud

Word cloud can be used to visualise the result of a text variable. They trim non essential word and allows to get a sense the frequency of redudant word within that variable.

Statistical Test

Two random variables are called independent if the probability distribution of one variable is not affected

by the presence of another. This is tested through the *Chi-squared Test of Independence* that allows to know if the independance hypothesis has a higher value than the .05 significance level.

The Chi-squared Test provide a *p-value*: if the p-value is greater than the .05, then the two tested variable are independent

Check for Errors

- **Sampling error**: The error that occurs when a sample of the population rather than the entire population is surveyed.
- **Coverage error**: The error that occurs when the sampling frame – i.e., the list from which the sample is taken – does not correspond to the population of interest. For instance, it might undercount migrants.
- **Unit nonresponse error**: The error when a designated respondent does not participate in the survey. A serious source of bias in the Literary Digest case.
- **Item nonresponse error**: The error when a respondent does not answer all of the questions, or answers "don't know".
- **Measurement error from Respondent**: Occurs when the respondent does not accurately answer the question.
- **Measurement error from Interviewer**: Occurs when interviewers do not pose the questions properly (or falsify the answers).
- **Postsurvey error**: The error that occurs in processing and analyzing survey data.

Discover

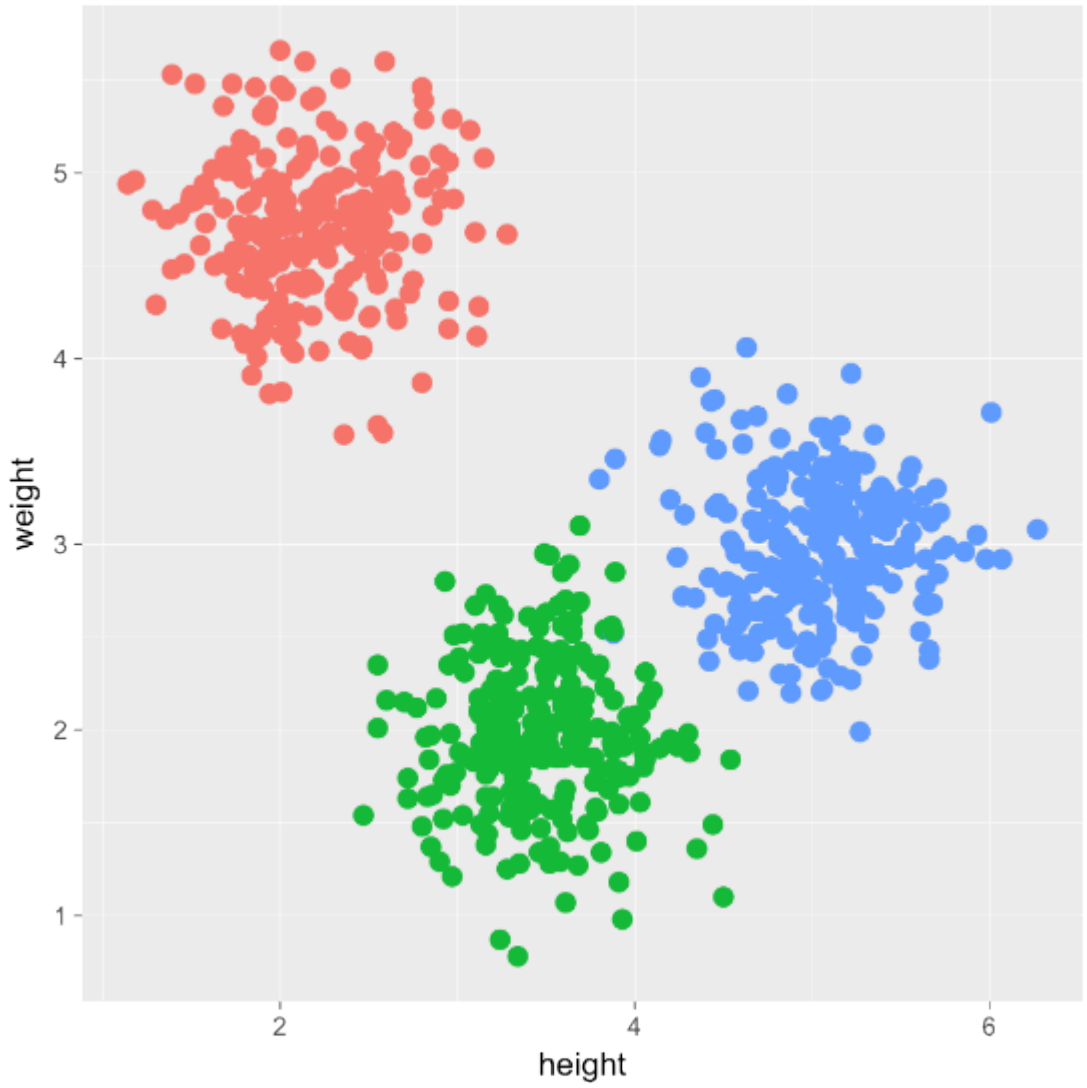
IMPORTANT

This chapter is not written yet.

Cluster analysis is a data-driven statistical technique that can draw out – and thence characterise – groups of respondents whose response profiles are similar to one another. The response profiles may serve to differentiate one group from another if they are somewhat distinct. This might be needed if the aim were, say, to define target groups for distinct safety net interventions. The analysis could help clarify the distinguishing features of the groups, their sizes, their distinctness or otherwise, and so on. Unfortunately there is no guarantee that groupings derived from data alone will make good sense in terms of profiling respondents.

Cluster analysis does not characterise the groupings; you have to study each cluster to see what they have in common. Nor does it prove that they constitute suitable target groups for meaningful development interventions Cluster analysis is thus an exploratory technique, which may help to screen a large mass of data, and prompt more thoughtful analysis by raising questions such as:

- Is there any sign that the respondents do fall into clear-cut sub-groups?
- How many groups do there seem to be, and how important are their separations?
- If there are distinct groups, what sorts of responses do “typical” group members give?



Multivariate Analysis

Refugees profile are defined by multiple categories. However it is very difficult for the human brain to process more than 7 categories toggether. An important challenge to understand the profile of the population is to discover how categories interact together. Fortunately, since the 70's, Social scientist have developed technique that allow to discover statistical clusters among a specific population.

Multiple Correspondence Analysis (MCA) is a data analysis technique for nominal categorical data, used to detect and represent underlying structures in a data set.

Dimensionnality reduction

The first step of the analysis is to reduce the numbers of dimansion in order to represent each observation in a 2D space.

Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Refugee data are mostly categorical so clustering is done on the result of the Multiple Correspondence Analysis.

Hierarchical Classification on Principle Components

Description of statistical clusters

Latent Class Models

it possible to conduct cluster analyses that are based on a statistical model, and not purely exploratory as are traditional cluster analysis techniques. The advantages of this approach to cluster analysis are that a number of diagnostic tools are available to help determine the appropriate number of clusters, and that the clustering is model-based in that it relies on a statistical model.

Predict

IMPORTANT

This chapter is not written yet.

Regression Analysis

Regression analysis is a statistical process for estimating the relationships among variables. Regression (and more specifically in the case of categorical data: logistic regression) can be used to predict certain characteristics or events linked to one household.

Analysis Type

The Multiple Linear Regression Model

The multiple linear regression model and its estimation using ordinary least squares (OLS) is doubtless the most widely used tool in econometrics. It allows to estimate the relation between a dependent variable and a set of explanatory variables minimizing the squared distances between the observed and the predicted dependent variable.

Key concepts: OLS Assumptions and Estimation, Goodness of Fit, Small Sample Properties, Tests in Small Samples, Confidence Intervals in Small Sample, Asymptotic Properties of the OLS Estimator, Asymptotic Tests, Confidence Intervals in Large Samples, Small Sample vs. Asymptotic Properties, More Known Issues (Non-linear Functional Form, Aggregate Regressors, Omitted Variables, Irrelevant Regressors, Reverse Causality, Measurement Error, Mutlicollinearity).

Functional Form in the Linear Model

Despite its name, the classical *linear* regression model, is not limited to a linear relationship between the dependent and the explanatory variables. It is indeed possible to build a model which is linear in the parameters but that also includes non-linear functions of the regressors.

Key concepts: Log-Linear, Semi-Log, Polynomial, Inverse, Dummy Variables, Interaction Terms, Spline Functions.

Heteroskedasticity in the Linear Model

This chapter relaxes the homoscedasticity assumption of the least squares estimation, and shows how the parameters of the linear model can be correctly estimated and tested when the error terms are heteroscedastic, i.e. their variance, conditioned on the regressors, changes across observations.

Key concepts: Groupwise Heteroskedasticity, Estimation with OLS, Estimating the Variance of the OLS Estimator, Testing for Heteroskedasticity, Estimation with GLS/WLS when the Variance Matrix is Known, Estimation with FGLS/FWLS when the Variance Matrix is Unknown.

Clustering in the Linear Model

This chapter relaxes the homoscedasticity assumption of the least squares estimation and allows the

error terms to be heteroscedastic and correlated within groups or so-called clusters. It shows in what situations the parameters of the linear model can be consistently estimated by OLS and how the standard errors need to be corrected. Clustering might arise when the sampling mechanism first draws a random sample of groups (e.g. schools, households, towns) and then surveys all (or a random sample of) observations within that group.

Key concepts: Random Cluster-Specific Effects, Estimation with OLS, Estimating Correct Standard Errors, Efficient Estimation with GLS, Estimating Correct Standard Errors with Random Cluster-Specific Effects.

Instrumental Variables

In many applications of the linear model, we suspect that some regressors are endogenous, i.e. one or more regressors are correlated with the error term. In this situation, OLS cannot consistently estimate the causal effect of the regressor on the dependent variable. Sometimes, we are able to find exogenous variables which are correlated with the endogenous regressor but not correlated with the error term. Such variables are called instrumental variables or instruments. If there are enough good such instruments, we can estimate the causal effect of the regressor on the dependent variable.

Key concepts: Canonical Examples (Omitted Variables, Simultaneity and Reversed Causality, Measurement Errors), Estimation with OLS, Estimation with IV (2SLS), Asymptotic Properties of the IV Estimators, What are Valid Instruments, Testing for Exogeneity of the Instruments, Testing for Relevance of the Instruments, Testing for Exogeneity of the Regressors.

Panel Data: Fixed and Random Effects

In panel data, individuals (persons, firms, cities, …) are observed at several points in time (days, years, before and after treatment, …). This chapter focuses on panels with relatively few time periods (small T) and many individuals (large N). This chapter introduces the two basic models for the analysis of panel data, the fixed effects model and the random effects model, and presents consistent estimators for these two models. Panel data are most useful when we suspect that the outcome variable depends on explanatory variables which are not observable but correlated with the observed explanatory variables. If such omitted variables are constant over time, panel data estimators allow to consistently estimate the effect of the observed explanatory variables.

Key concepts: The Random Effects Model, The Fixed Effects Model, Estimation with Pooled OLS, Random Effects Estimation, Fixed Effects Estimation, Leas Squared Dummy Variable Estimation (LSDV), First Difference Estimator, Time Fixed Effects, Random Effects vs. Fixed Effects Estimation.

Binary Response Models

Many dependent variables of interest in economics and other social sciences can only take two values. The two possible outcomes are usually denoted by 0 and 1. Such variables are called dummy variables or dichotomous variables. As already seen in the course *Introductory Econometrics*, there are several ways to model these outcomes using regressions. This chapter specifically focuses on the interpretation of the Probit and Logit models, and of their estimated parameters.

Key concepts: The Econometric Model: Probit and Logit, Latent Variable Model, Interpretation of the Parameters, Estimation with Maximum Likelihood, Estimation with OLS.

Limited Dependent Variable Models

- The effect of *truncation* occurs when the observed data in the sample are only drawn from a subset of a larger population. The sampling of the subset is based on the value of the dependent variable.
- *Censoring* occurs when the values of the dependent variable are restricted to a range of values. As in the case of truncation the dependent variable is only observed for a subsample. However, there is information (the independent variables) about the whole sample.
- The *sample selection problem* occurs when the observed sample is not a random sample but systematically chosen from the population. Truncation and censoring are special cases of sample selection or incidental truncation.

This chapter presents the econometric models that are used to deal with the above-mentioned situations.

Key concepts: Truncation, Truncated Regression, Interpretation of Parameters, Estimation; Censoring, Tobit Model Type I, Interpretation of Parameters, Estimation; Selection, Heckman Selection Model, Interpretation of Parameters, Estimation, Estimation with Maximum Likelihood, Estimation with Heckman's Two-Step Procedure.

Quantile Regression

Quantile regression provides an alternative to ordinary least squares (OLS) regression and related methods, which typically assume that associations between independent and dependent variables are the same at all levels. Quantile methods allow the analyst to relax the common regression slope assumption. In OLS regression, the goal is to minimize the distances between the values predicted by the regression line and the observed values. In contrast, quantile regression differentially weights the distances between the values predicted by the regression line and the observed values, then tries to minimize the weighted distances.
Two empirical applications are here attached to understand why and when it may be appropriate to use

this model:

- Thinking beyond the mean: a practical guide for using quantile regression methods for health services research
- A gentle introduction to quantile regression for ecologists

Advanced Time Series Analysis

Additionally to the forecasting models presented in the chapter *Introduction to Time Series Regression and Forecasting* of the previous course, there are other time series regressions used for forecasting which involve lags of both the dependent variable and the error term, so called AR(I)MA. Before introducing these models, the concept of stationarity and the technique of differencing time series are discussed. The course also includes applications in R.

Key concepts: Stationarity and Differencing, Autoregressive Models, Moving Average Models, Non-seasonal AR(I)MA Models, Estimation and Order Selection, AR(I)MA Modelling in R, Forecasting, Seasonal AR(I)MA Model.

Multiple Hypothesis Testing

When a set of hypotheses are tested simultaneously and independently from each other, then the probability of rejecting at least one of the true null hypothesis can become excessively high. Methods for dealing with multiple testing frequently call for adjusting the significance level in some way, so that the probability of observing at least one significant result due to chance remains below your desired significance level.

Key concepts: the Problem of Multiple Testing, Bonferroni Correction, False Discovery Rate, Comparison of the Correction Methods.

For a further understanding of this topic, have a look at this interactive reading.

Advise

IMPORTANT

This chapter is not written yet.

Composite indicators

- A way to compile together multiple indicators
- Allow to capture the complexity of a situation
- Implies to define a calculation method

Index for locations

Index for Individuals/households

- Headline money metric measures of poverty has limitations
- If someone is deprived in a third or more of ten (weighted) indicators, the index identifies them as 'poor', and the extent – or intensity – of their poverty is measured by the number of deprivations they are experiencing.

Explore indicator correlations

- Not too high,
- not too low…

Decide how to aggregate indicators

- Indicators that can be added together : Vouchers + Cash
- Indicators that should be multiplied (no compensability): Access to education & Access to health care

Assess robustness & sensitivity analysis

Ensure that the decision made for normalization, weighting and aggregation have limited effects on the final ranking

Data Crunching

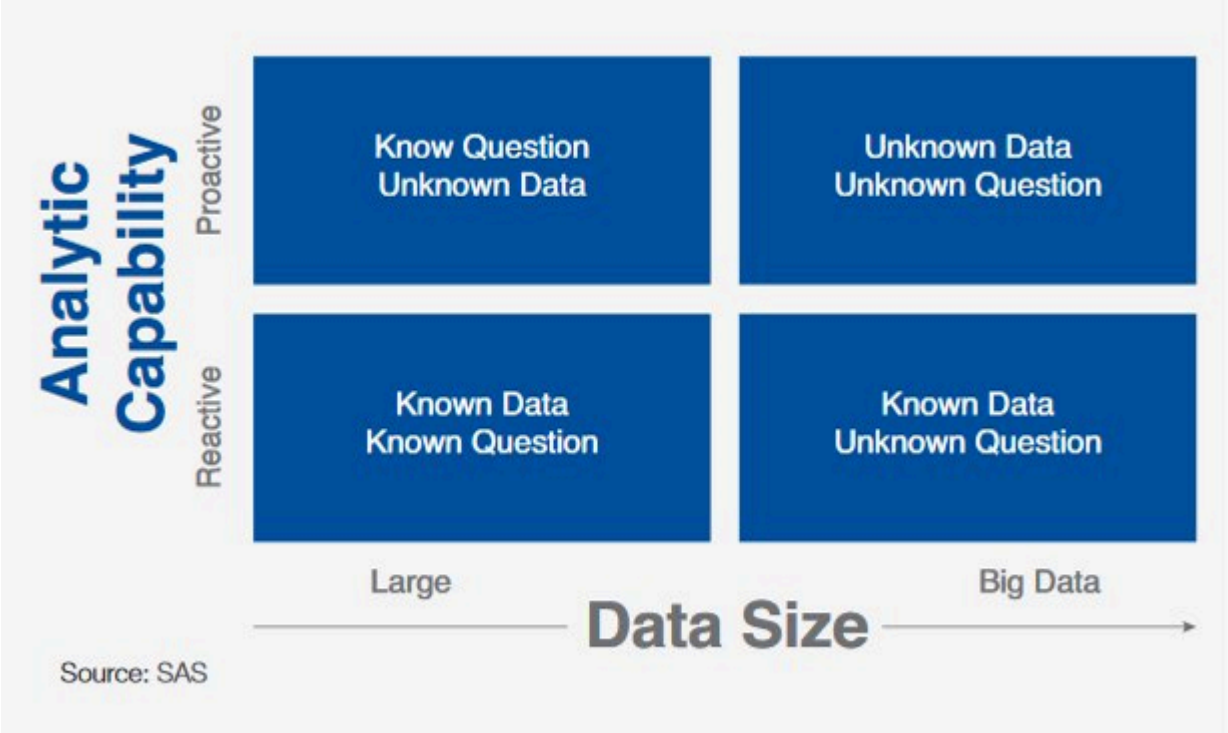
NOTE

Key Take Away :

Challenges with Household Survey analysis

Household survey often results in dataset with over 300 variables to process & explore. Deadline to get insight from the dataset are often tight and Manual processing is very lengthy and can be done only for a limited part of the dataset. Often, because of those challenges, a lot of potential insights are not discovered.

To address this, it's necessary to move from a reactive support to a proactive one.



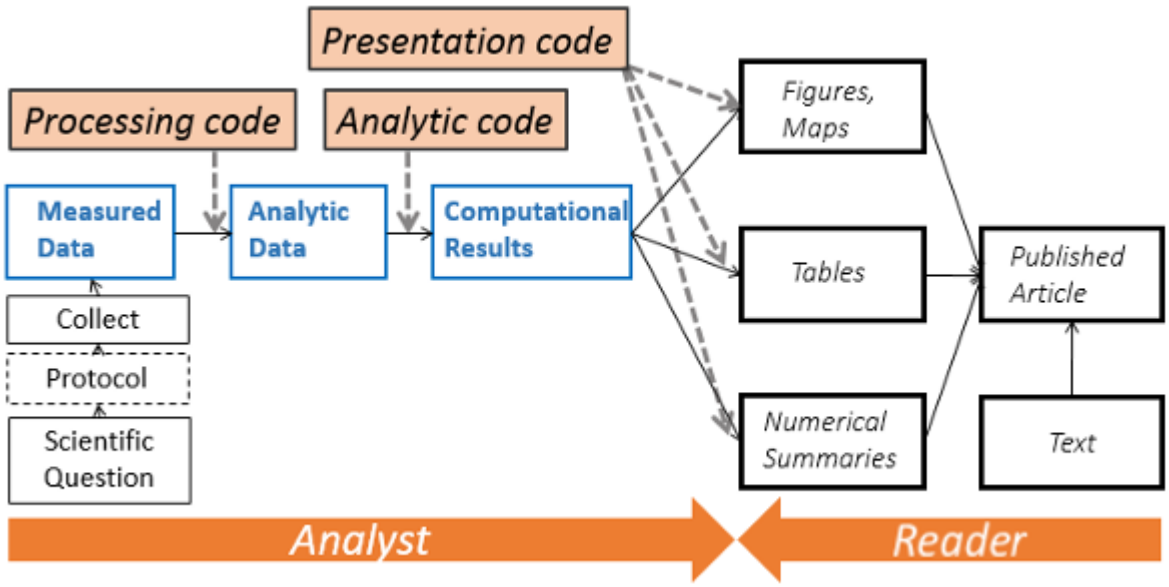
Data Crunching is about automating processes:

- Upstream process of data interpretation: consists of the **preparation of a dataset** so that it can be processed, sorted and structured to run algorithms and program sequences on it;
- Can **save a lot of time** as the processes do not need to be performed manually (different from data munging and data wrangling that refers manual processing of data);
- Can be **iterative** when the output of the crunching process includes new data or reveals errors. This means that the program sequences may be repeated until the desired result is achieved;
- Enable better **analysis reproducibility** (minimize point & click interventions) to facilitate peer review

Data Crunching Automation

Using the right combination of packages from the R statistical language, it is possible to integrate all necessary data analysis steps into **scripts**:

- Data management (clean, recode, merge, reshape)
- Data analysis (test, regression, multivariate analysis, etc…)
- Data visualisation (plot, map, graph…)
- Writing up results (report and presentation generation)



Using KoboloadeR to facilitate reproducibily.

Before data visualization & interpretation many steps are required:

Import raw data

In a structured survey with numbered questions, the **flat file** type of data structure has a column for each question, and a row for each respondent. A more complex survey data structure arises if the data are **hierarchical**. A common type of hierarchy is where a series of questions is repeated say for each child in the household, and combined with a household questionnaire, and maybe data collected at community level. For analysis, we can create a rectangular flat file, at the 'child level', by repeating relevant household information in separate rows for each child. Similarly, we can summarise information for the children in a household, to create a 'household level' analysis file.

In the case of hierearchical dataset, it is required to use ODK Briefcase to export and configure the key to join the frame together.

Recode & Relabel

- Leverage the same xlsform file (saved as .xls – not .xlsx) already used to encode the questionnaire to generate a data dictionary
- Extend xlsform by adding additional column (chapter, disaggregation, correlation, etc.)

- Potentially revise label wording to make them more concise when they will appear on the output chart

Clean records

- Cleaning Log defined though the iteration of the crunching
- Log stored as a worksheet, act as documented data audit trail
- Log actions to be sorted as “update” or “delete”
- Log to be re-applied every time to raw date

Reweight dataset according to sampling strategy,

- Associate a weight to each record
- Weight defined by the sampling script (can be based on cluster, or Respondent Driven Sample)
- Possibility to use post-stratification to re-compute corrected weights in case of low coverage of the sample

Build new indicators from existing variable, etc.

Indicators are summary measures. They often provide a baseline from which to weigh up the finer points. It is important not to create unnecessary confusion. An indicator should synthesise information and serve to represent a reasonable measure of some issue or concept. The concept should have an agreed name so that users can discuss it meaningfully.

- Create new indicators from existing one: Need to define in a worksheet for each indicator: type, name, label, chapter, correlation, aggregation, formula, frame
- Indicators formula written with a R-ready syntax: Allow for complex notation:
    - May need to use dcast if you want to calculate an indicator based on values from a nested data frame. dcast will work as pivot table using the unique ID used for the join.
    - May need to use if when trying to do a calculation where you could have potential zero as numerator Indicators are calculated, appended to the right data frame and then the indicator definition is appended in the data dictionary

KoboloadeR

It is An R packages (i.e. a series of functions) that can be plugged to a configuration file in order to separate “input”, “processing”, and “output”

- The “output” is an Rmd (Rmarkdown) file than produce word, pdf or html reports

- The configuration file includes references to all “input”:

```
Path to raw data files
Path to form (in xlsform) in order to build a data dictionary
Path to the sample weight
Path to the data cleaning log
Path to the indicator calculation sheet
```

Collaborative Development

- Open Source Package maintained in GitHub
- Submit issues for bug report or feature request in Github
- Fork and submit pull request for code review and integration

Analysis Workshop

IMPORTANT

The Goal of the analysis workshop is to review of data patterns in order to collect qualitative interpretations. This workshop should stimulate discussions from each presentation. input provided and any decisions taken about changes or additions to the analysis presented should be cautiously recorded.

Workshop preparation: identify patterns

While identifying categories and sorting the data, you will start to see patterns within and between categories. At this stage it will be important to assess the relative importance and relationships between those categories. You may want to look at:

- How many times a specific category came up?
- What are the key ideas being expressed?
- What similarities and differences are being expressed?
- If two or more categories consistently appear together for one group?

Relationships might suggest a cause and effect; this will help you to explain why something occurs. You can ask yourself:

- How do things relate?
- What data supports this interpretation?
- What responses are contradicting findings and/or do not fit into categories?
- Are they important to the overall understanding?

Data Digest Presentation

Data Analysis Workshop are among common good practices. .

- Engage with various levels (Field, Sector, Partner, Authority..)
- Facilitation around “data digest” (±60 slides for 3 hours session)
- Rapporteur to capture (extensive note taking):

Faciliation & Note taking

Having sorted and identified the relationships, relative importance and possible causes, you need to bring it all together. You have to interpret the data: what does it all mean?

- Take a step back and look at what you learned. Ask yourself for example:
- What will those who are using the findings of the information be most interested in?
- What new things did you learn about the protection situation?
- What main protection risks for several and/or specific population groups did appear?
- What is the scale of a given problem/protection risks?
- Are there particular groups within the population who are particularly exposed to specific problems/risks?
- Did you confirm patterns mentioned in earlier reports or did you see new trends?
- The protection problems observed are there isolated incidents or generalizable trends?
- Are the incidents related to the crisis/emergency/conflict or are they more related to endemic issues such as cultural practices or generalized poverty?

Practical steps during the slides review

- **Reflect**: 

Data quality and or suggestions to change questions

- **Interpret**: 

Qualitative interpretations of data patterns

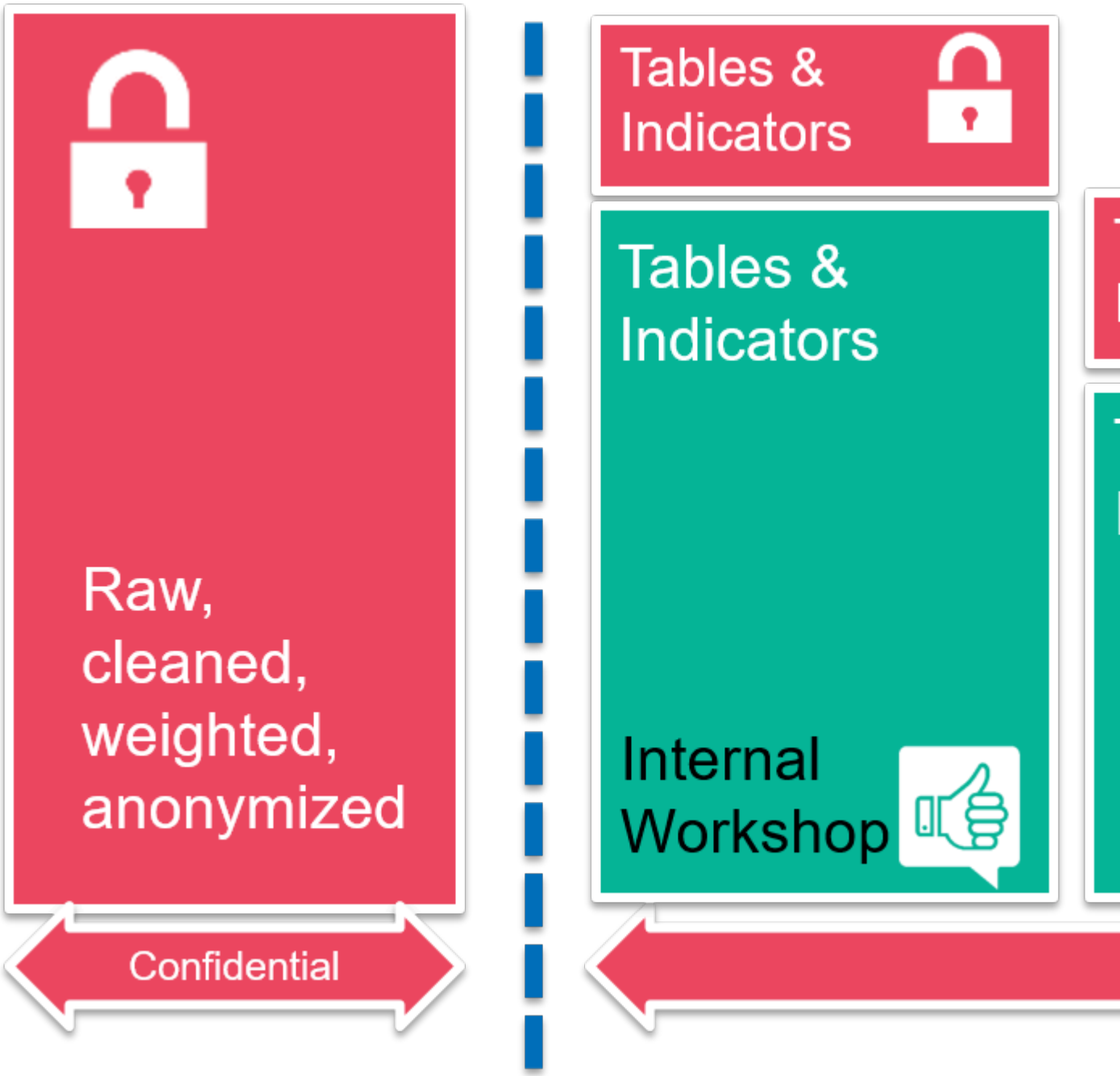
- **Recommend**: 

Programmatic adjustment (3RP or COP design)

- **Classify**: 

Level of sensitivity for the information

Clearance process



Styling guide

Confidence intervals:

- Different software often calculate confidence intervals as being negative and hence below zero or above 100. Negative confidence intervals or CI above 100 are meaningless. In the report, always round negative confidence intervals to '0' and round those above 100 to '100'.

Rounding decimal points:

- Make sure to round properly decimal points according to basic rules:
- When decimal is between 1-4, round down.
- When decimal is between 5-9, round up.

Decimal points in the results:

- When the results is a whole number e.g. 30%, make sure to always write 30.0% with the '.0' in the decimal place in the report. This ensures that the decimal point was not forgotten and is actually equal to zero.

Missing data or consent not provided:

- Data should be excluded from all analysis and should not be accounted for in the denominator.

Always clean the data first before going into analysis:

- Frequencies and means should be run on categorical and continuous variables, respectively.
- Missing data should be looked at and a record of them should be kept.

Age variable:

- When selecting age or creating an age variable category in Epi info software from the 'months' variable generated by ENA, don't forget the '.99' otherwise some children with an exact birth

date may be excluded from the analysis. E.g. 6-23.99 (and not 6-23 or 6-23.9).

Reproducible Analysis

Always save newly generated variables into a new data file named following a naming convention to be respected by all involved in the survey data analysis.

Model for Final Report

IMPORTANT

This chapter is not written yet.

Slides & Infographics

IMPORTANT

This chapter is not written yet.

Experience shows that it is easier for decision makers to see trends in data when it is visually presented in charts, maps and graphs. While valuable time should not be wasted making a visually flawless final report, attracting the readers' eyes to key messages and actionable information is strategic.

Organise Microdata for Social Scientist

IMPORTANT

Data Confidentiality, Data Security and Data sensitivity are two important consideration but should not be confused.

- Data *Confidentiality* is linked to data protection and can be addressed through anonymisation.
- Data *Security* is dependant from technical processes that needs to be established to prevent leaks.
- Data *Sensitivity* is tied a collective clearance and information classification process.

Once those elements are addressed, it becomes possible to engage with researchers.

Dealing with confidentiality

Once anonymised, a dataset does not fall anymore under the Policy on the Protection of Personal Data.

Anonymization techniques

Even when personal data is not being collected it still may be appropriate to apply the methodology since quasi-identifiable data or other sensitive data could lead to personal identification or should not be shared.

Type	Description
Direct identifiers	Can be directly used to identify an individual. E.g. Name, Address, Date of birth, Telephone number, GPS location
Quasi-identifiers	Can be used to identify individuals when it is joined with other information. E.g. Age, Salary, Next of kin, School name, Place of work
Sensitive information	& Community identifiable information Might not identify an individual but could put an individual or group at risk. E.g. Gender, Ethnicity, Religious belief
Meta data	Data about who, where and how the data is collected is often stored separately to the main data and can be used identify individuals

The following are different generic anonymisation actions that can be performed on sensitive fields. The type of anonymisation should be dictated by the desired use of the data. A good approach to follow is to start from the minimum data required, and then to identify if any of those fields should be obscured.

The methods below can be referenced in the dedicated column within xlsform (cf above)

Type	Description
Remove	Variable is removed entirely from the data set. The Variable is preserved in the original file.
Reference	Variable is removed entirely from the data set and is copied into a reference file. A random unique identifier field is added to the reference file and the data set so that they can be joined together in future. The reference file is never shared and the Variable is also preserved in the original file.
Mask	The Variable values are replaced with meaningless values but the categories are preserved. A reference file is created to link the original value with the meaningless value. Typically applied to categorical Variable . For example, Town names could be masked with random combinations of letters. It would still be possible to perform statisitical analysis on the Variable but the person running the analysis would not be able to identify the original values, they would only become meaningful when replaced with the original values. The reference file is never shared and the data is also preserved in the original file.
Generalise	Continuous Variable is turned into categorical or ordinal Variable by summarising it into ranges. For example, Age could be turned into age ranges, Weight could be turned into ranges. It can also apply to categorical Variable where parent groups are created. For example, illness is grouped into illness type. Generalised Variable can also be masked for extra anonymisation. The Variable is preserved in the original file.

Statistical disclosure control (SDC)

Though there's a few articles about the failure of anonymization that shows how removing names & ID is not always sufficient to prevent "data re-identification".

Many techniques can be used for "statistical disclosure control": suppression, inference control, banardisation, rounding or sampling. Other approaches includes rules like for instance "do not share figures for a spatial unit if it does not reach the 1000 refugees threshold"…

A dedicated R module is available to perform anonymisation analysis.

Ensuring data security

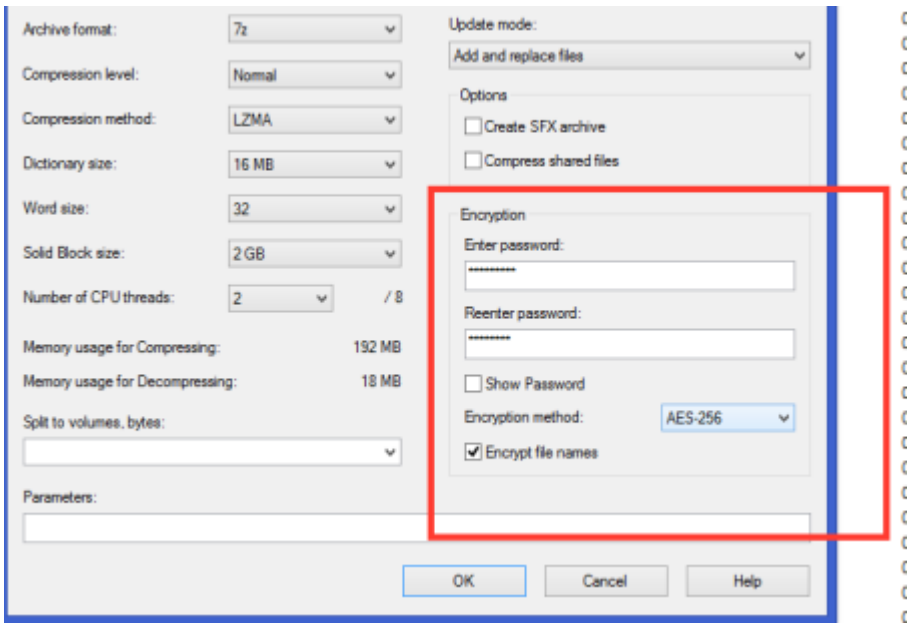
Access Registry

A first requirement is to set up a standard registry of person who work on UNHCR datasets. This is actually prescribed in the data protection policy.

Sharing via a safe mechanism: File encryption

What is a safe mechanism to share information: for instance which software to use for encryption, how to share password, etc. Potential requirements could include: - Use a well know encryption approach – The common standard si AES -Advanced Encryption Standard (AES) - Rely on open source software – so both parties can easily encrypt & decrypt without being tied to software procurement obstacle. - Combine encryption and file compression: so files are easier & lighter to share - The password used for the encryption should be at least 10 character long with a mixture of lowercase and uppercase alphabetic character, numbers and symbols. This should allow to build what is commonly called a strong password and should always be transmitted independently form the file (for instance on a separate paper sheet with no reference to the file it allows to open).

In terms of software, it is possible to use 7zip.



A summary of the principle above woud be:

Data files should be encrypted with AES-256 method using a strong password (at least 10 character long with a mixture of lowercase and uppercase alphabetic character, numbers and symbols) and compressed using the 7zip format with the 7zip software. Password will be transmitted printed on a paper that will need to be secured by the receiving agency.

Dealing with sensitive information

Information classification

Sensitive Data - institutional data that is not legally protected, but should not be made public and should only be disclosed under limited circumstances. Users must be granted specific authorization to access since the data's unauthorized disclosure, alteration, or destruction may cause perceivable damage to the institution.

Data sharing for research

If outsourced, formal agreement needs to established.

> The UNHCR and Partner Name will identify the staff to be part of the joint research team. Any data shared under this agreement will not be provided to any third party. For its part, UNHCR agrees to share defined and agreed upon data with the Partner Name for the purposes of the Partner Name and UNHCR collaboration on this project herein-defined as "Project Name". All information that would allow for identification of individuals will be excluded from these datasets, e.g. refugee ID number. UNHCR will share this information via a safe mechanism to reduce the likelihood of a third party accessing the data unlawfully. Partner Name will specify by name and title who will receive the information, who will have access to the information, and where the information will be kept, e.g. individual personal computer or server, all with the intent to avoid unlawful access and use of the information. Once the information is used for its defined purpose, the data will be disposed of at

> a date determined and in agreement by the two parties.

Restricting publication of findings

Research Confidentiality agreement are written and legally-binding Confidentiality Agreement that must be signed by the lead researcher, all members of the research team that will have access to individually identifiable information from the records. The agreement coudl include the following points:

> *Analysis Project Title* *Principal Investigator:* *UNHCR*
>
> *I,* *Resesarcher Name*, *from* *Resesarch Organisation Name*, *as a member of this research team, understand that I may have access to confidential information about study sites and participants. By signing this statement, I am indicating my understanding of my responsibilities to maintain confidentiality and agree to the following:*
>
> 1. *keep all the research information shared with me confidential by not discussing or sharing the research information in any form or format (e.g., disks, tapes, transcripts) with anyone other than the Researcher(s).*
> 2. *keep all research information in any form or format (e.g., disks, tapes, transcripts) secure while it is in my possession.*
> 3. *return all research information in any form or format (e.g., disks, tapes, transcripts) to the Researcher(s) when I have completed the research tasks.*
> 4. *after consulting with the Researcher(s), erase or destroy all research information in any form or format regarding this research project that is not returnable to the Researcher(s) (e.g., information stored on computer hard drive).*
> 5. *notify the local principal investigator immediately should I become aware of an actual breach of confidentiality or a situation which could potentially result in a breach, whether this be on my part or on the part of another person.*

Engaging in research

Reproducible research

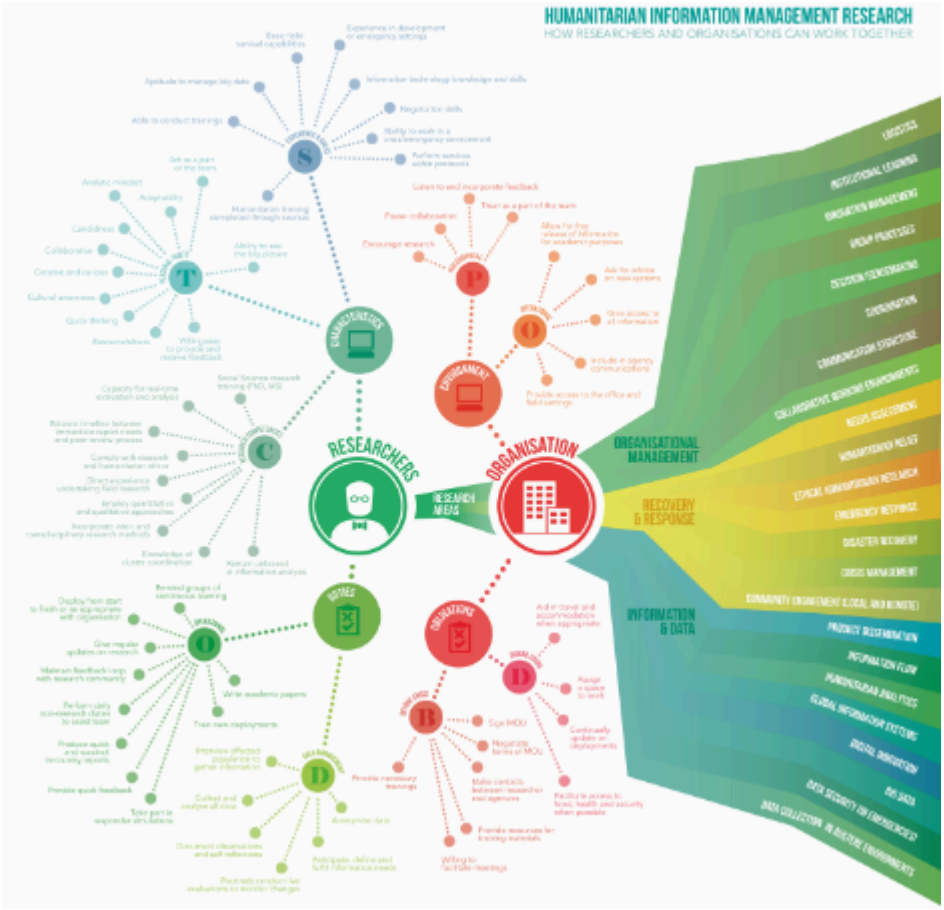
To ensure that research done on the dataset can be reproduced afterwards by internal staff both to check them and to refresh the analysis when we have new data a series of good practices shoudl be implemented:

1. For every result, **keep track** of how it was produced
2. **Avoid manual data manipulation** steps
3. **Archive** the exact versions of all external programs used
4. **Version control** all custom scripts
5. **Record all intermediate results**, when possible in standardized formats
6. For analyses that include randomness, **note underlying random seeds**
7. Always **store raw data** behind plots
8. Generate hierarchical analysis output, allowing layers of increasing detail to be inspected
9. Connect **textual statements** to underlying results
10. Provide **public access** to scripts, runs, and results

Establish a survey catalog

Humanitarian Research in the context of social science and data analysis is still new but can benefit the organisedtion for instance to:

- Co-development and co-design of tools, protocols, products, processes, and innovations
- Facilitate organisational learning, keeping track of lessons learned, and providing a neutral stance for moderating innovation and change processes
- Access to wider body of knowledge, from academia or other organisations, and research in other fields.



To facilitate this process, the first approach woudl be to document the dataset according to the Data Documentation Initiative (DDI) metadata standard developped by the International Household Survey Network (IHSN).

Once the metadata are generated in the right format, it becomes possible to publish them within the ISHN Microdata catalog or the World Bank Microdata Library

Open Data

IMPORTANT

This chapter is not written yet.