

Research Design Guidance: Sampling

April 2020



REACH Inventing more effective humanitarian

IMPACT Shaping practices
Influencing policies
Impacting lives

Cover photo: Enumerator collecting settlement information in Novoluhanske, Ukraine © IMPACT Initiatives




About IMPACT

IMPACT Initiatives is a Geneva based think-and-do-tank, created in 2010. IMPACT is a member of the ACTED Group. IMPACT's teams implement assessment, monitoring & evaluation and organisational capacity-building programmes in direct partnership with aid actors or through its inter-agency initiatives, REACH and Agora. Headquartered in Geneva, IMPACT has an established field presence in over 20 countries. IMPACT's team is composed of over 300 staff, including 60 full-time international experts, as well as a roster of consultants, who are currently implementing programmes across Africa, Middle East and North Africa, Central and South-East Asia, and Eastern Europe.

Introduction

Sampling is the **process of selecting units (i.e. a sample) within the wider population of interest**, so as to be able to make inferences and estimate characteristics and behaviour of the wider population. Sampling is **different from census** which is when every single unit within the wider population of interest is covered for the research. A complete census is often not practical or possible in a humanitarian or development setting.¹

In general, the key advantages of sampling are:

 <p>To reduce cost</p>	<p>It is obviously less costly to obtain the required information from a selected subset of a population (sample), rather than the entire population (census)</p>
 <p>To speed up data collection and analysis</p>	<p>Observations are faster to collect and summarize with a sample than with a census, simply because of the smaller number of observations required.</p>
 <p>To enhance the scope of the assessment</p>	<p>Sampling enables you to increase coverage of the study (for e.g. include more geographical areas or a wider population of interest) as well as to enhance the quality of the data being collected, in comparison to a census. Additionally, because of the lesser number of observations to be collected for a sample, highly trained personnel or specialized equipment, which are often limited in availability, can be used.</p>

This document provides practical guidance on how IMPACT country teams should go about this crucial step of sampling during the research design stage of any research cycle. Specifically, it will cover:

- Overview of probability and non-probability sampling;
- Types and applicability of different probability and non-probability sampling strategies;
- How to operationalise the selected sampling strategy;
- Frequently asked Questions (FAQs) related to sampling, including methodological issues encountered during data collection.

It is important to note that this document is **not aiming to provide a full “textbook” guide on everything related to sampling** or an exhaustive overview of all the different sampling strategies currently out there. On the contrary, the aim is to **outline the key aspects of sampling and the different types (and combinations) of sampling strategies that are specifically important to know about when implementing any research cycle at IMPACT.**

This Guidance Note is part of the wider (forthcoming) IMPACT Research Design Guidelines which will include practical guidance for all other steps of the research design process. The Research Design Guidelines is aimed to be finalised and released by Q2 2020.

¹ One exception is assessments of camps and sites, where a census assessment can be done relatively quickly (the surface area where the population lives is limited) and can have the combined purpose of providing a population count.

Table of Contents

Introduction	2
1. Types of sampling	5
1.1 Probability sampling	5
1.2 Non-probability sampling	7
1.3 Generalizability based on sampling type	7
2. Key parameters for sampling	9
2.1 Define geographical area and population of interest	9
2.2 Define unit of measurement.....	9
3. Types and applicability of different probability / non-probability sampling strategies	11
4. Operationalising the selected sampling strategy	19
4.1 Prepare sampling frame	19
4.2 Calculate sample size.....	19
4.3 Finalise strategy to select units within sampling frame i.e. identify participants for data collection ...	20
5. Frequently asked Questions (FAQs)	22
5.1 FAQs on choice of sampling strategies	22
5.2 FAQs on operationalising sampling strategies	23
5.3 FAQs on mitigating methodological issues encountered during data collection	27
6. Annexes	28
Annex 1: Memo on cluster sampling.....	28
Annex 2: User Guide for IMPACT’s online sampling tool	28
Annex 3: Example from REACH Jordan- A guide to GIS-based sampling for host community projects .	32
Annex 4: Sampling considerations for remote, phone-based data collection	41
Annex 5: Troubleshooting issues encountered due to inaccurate information in sampling frame	42
Annex 6: Additional reading materials	48

Figures and Tables

Figure 1: Generalisability based on sampling type.....	8
Figure 2: Depth of information based on unit of measurement	10
Figure 3: Overview of the types of probability and non-probability sampling strategies.....	11
Figure 4: Decision tree for choosing an appropriate probability sampling strategy:	18
Figure 5: Systematic selection on site – Option 2	25
Figure 6: Decision tree in case of uncertain access.....	47
Table 1: Checklist for the selection of geographical areas during research design	9
Table 2: Types and applicability of different sampling strategies	12
Table 3: Example sampling frame for refugee households in Jordan, stratified by region and time of arrival.....	19
Table 4: Example saturation grid for data collection using non-probability sampling	20
Table 5: Strategies for random selection of research participants – probability sampling.....	21

1. Types of sampling

There are two types of sampling: (a) probability sampling and (2) non-probability sampling.

1.1 Probability sampling

A sampling strategy in which a sample from a larger population is chosen in a manner that enables findings to be generalized to the larger population²

- The most important requirement for probability sampling is the random selection of respondents. This does not mean randomly interviewing households or individuals on the street. Instead, random selection means that each unit within the population of interest has an *equal probability* of being selected for the study, with the probability of selection being inverse to the population size (i.e. 1/ population size). This randomization ensures that a probability sample is representative and can be generalized to a population with a known level of statistical precision.³
- It is therefore important to mitigate any biases in the selection of a probability sample. Any bias reducing or even eliminating the probability of being selected amongst certain units means the sample can no longer be considered truly representative of that portion of the population.
- The second key requirement for probability sampling is to use statistical theory to calculate the minimum required sample size⁴ i.e. to calculate the required size of a probability sample (e.g. number of household or individual surveys to be conducted) based on the target level of statistical precision required for the research findings. Inferences of statistical precision are based on:
 - **Confidence level** i.e. the probability that the observed value of a parameter falls within a specified range of values
 - This is expressed as a percentage and represents how often the sample observation is truly generalizable; in other words, how often a true percentage of the population would pick an answer as provided by the sample.
 - For example, a study was conducted on a population of 300 million households with a sample size of 2,000 to generate findings generalizable with a 95% level of confidence. One of the key findings from this study could be “38% of the assessed population (sample) state that their health insurance coverage has changed over the past year”. With a population of 300 million, it is impossible to know exactly how many people would actually say yes to this, without conducting a full census. However, probability sampling with a 95% confidence level enables the researcher to make the best possible guess. Here, the 95% confidence level is telling us that if the survey was to be repeated over and over again, the results would match the answers from the actual population, within a specified range of values, 95% of the time.⁵
 - Therefore, the higher the confidence level, the more robust the study will be. While 95% is most commonly used, it is within acceptable standards to have a confidence level ranging between 90-99%: However, within IMPACT, we prefer not to go below a confidence level of 95%. A 100% confidence level does not exist as it implies a census.

² This type of generalisation is possible due to Probability Theory discoveries, in particular of the *Central Limit Theorem*, which can be traced back as early as 1733 (Salkind, 2010, Encyclopedia of Research Design).

³ Creswell, John W.; 'Research Design: Qualitative, Quantitative and Mixed Methods Approaches' (Third Edition, 2009); p.148

⁴ The formula used within IMPACT/ REACH for calculating sample size for probability sampling was first outlined by Krejcie and Morgan in 1970. The formula is $n = \frac{\chi^2 \cdot N \cdot p(1-p)}{\beta^2 \cdot (N-1) + (\chi^2 \cdot p(1-p))}$; where n=sample size, χ^2 = Chi-square for the specified confidence level at 1 degree of freedom, N=Population size, P= Population proportion (assumed to be 0.5 to generate maximum sample size), β = desired Margin of Error (expressed as proportion)

⁵ Adapted from <https://www.statisticshowto.datasciencecentral.com/confidence-level/>

- **Confidence interval / margin of error** i.e. an estimate in probability sampling of the range of upper and lower statistical values (+/-) that are consistent with the observed data and are likely to contain the actual population mean or percentage.⁶
 - This is expressed as +/- to estimate the spread of the mean or percentage for which we are likely to estimate properly the population mean or percentage for a certain characteristic based on observations in the sample.⁷
 - For example, if the findings from a household survey was that 50% of the households were found to be living in inadequate shelters, the inference from a sample studied with a +/- 5% margin of error would be that we can expect the results for the entire population to be roughly between 45% to 55%.⁸ If the study used a 95% confidence level, we can conclude that 95% of the time, we can expect the results for the entire population for this particular occurrence to be between 45% to 55%.
 - Therefore, the narrower the confidence interval, the more robust the study.
- **[For experimental survey design]⁹ Statistical power** i.e. an estimate of the probability of making a type II error which is wrongly failing to reject the null hypothesis for a binary hypothesis test. In other words, statistical power is an estimate of the probability of accepting the alternative to the null hypothesis, when the alternative hypothesis is true i.e. the ability of a test to detect a specific effect within the observed sample, if that specific effect exists in reality.¹⁰
 - Statistical power is expressed numerically between a range of 0 to 1.
 - A sample size with 95% confidence level and 5% margin of error assumes a statistical power of 0.8. This also means that with this sample size, 20% of the time, we are likely to be making a type II error.
 - As the statistical power increases, the probability of making a type II error decreases. As such, the higher the statistical power factored in, the more robust the study is likely to be.
 - For example, we have to conduct an endline evaluation of a USAID project in Jordan. Sample sizes were calculated to produce results with a confidence level of 95% and with a statistical power of 0.8, assuming a difference in proportion between groups of at least 10%. What does this mean? It means:
 - The anticipated effect of the USAID interventions was to bring about a 10% change in the proportion of households that experience a specific outcome (for e.g. low food consumption scores) over the course of the project (let's say five years)¹¹
 - The statistical power of 0.8 ensures a relatively low chance of identifying that no impact or change in outcome is detected during the endline analysis, when in fact there has an impact.
- In sum, the key purpose of employing probability sampling is to enable the researcher to generalize from a sample to a wider population of interest so that inferences can be made about some characteristic, attitude or behavior, based on the trends observed within the sample.¹²

⁶ Creswell, John W.; 'Research Design: Qualitative, Quantitative and Mixed Methods Approaches' (Third Edition, 2009); p.228

⁷ Adapted from <https://www.statisticshowto.datasciencecentral.com/confidence-level/>

⁸ It is worth noting that sampling calculators usually assume findings of 50% because of which the margin of error/ confidence interval actually shrinks the further you get from a 50% finding. So depending on how the finding is +/- 50%, the actual margin of error will be less than +/- 5%.

⁹ This is a research approach where independent variable(s) are manipulated and applied to dependent variables to measure their impact on the latter. Since one of the primary purposes of such an experimental research design is to detect an effect, it is important to factor in statistical power into the research design.

¹⁰ The following online tool can be used to calculate sample sizes with statistical power parameters: <https://clincalc.com/stats/samplesize.aspx>

¹¹ This is also known as the effect size i.e. an estimate in probability sampling that identifies the strength of the conclusions about group differences or the relationships among variables in quantitative studies.

¹² Babbie, E.; 'Survey Research Methods' (Second Edition, 1990)

1.2 Non-probability sampling

A sampling strategy in which a sample from a larger population is chosen purposefully, either based on (1) pre-defined selection criteria based on the research questions and objectives or (2) a snowball approach to build a network of participants from one entry point in the population of interest.

- Although not generalizable with a known level of statistical precision, non-probability sampling can still generate indicative findings with some level of representation if the targeting of participants is done correctly. A standard good practice in this regard is to develop a list of potential respondent types or profiles, based on the objectives of the research. For example, if we are conducting an assessment of the education needs of refugee children in a specific context, we could consider children of school-going age or their caregivers, teachers or staff at schools in the areas, and aid actors working on education.¹³
- Sample sizes for non-probability sampling are based on what is feasible and what should be the minimum to meet the research objectives with quality standards.¹⁴
 - One of the key guiding principles to determine sample sizes for non-probability sampling is to lead sampling by saturation i.e. continue conducting interviews and discussions until data saturation has been achieved and no new themes or issues are appearing in the data collected.
 - Alternatively, quotas or thresholds can be set based on what is known about the population of interest; for e.g. if we want to conduct FGDs to understand a population's ability to access basic services across three different districts, it would make sense to: (1) conduct a minimum of two FGDs per district, one male and one female; and (2) conduct two additional FGDs in District 2 because it also has a large internally displaced population whose experiences may be different from the overall population.
- Non-probability sampling is often used as an alternative to probability sampling when this is unfeasible, often due to time, access or resource limitations. Given the requirement for probability sampling to have possible access to *every unit* in the population of interest, it is sometimes almost impossible to work with it in contexts where security or other limitations disrupts access, or in contexts where there is very little known about the population of interest.
- Therefore, the key difference between probability and non-probability sampling is that with probability sampling, if done correctly, the data and findings can be considered representative of and generalizable to the wider population being studied with a known level of statistical precision.

1.3 Generalizability based on sampling type

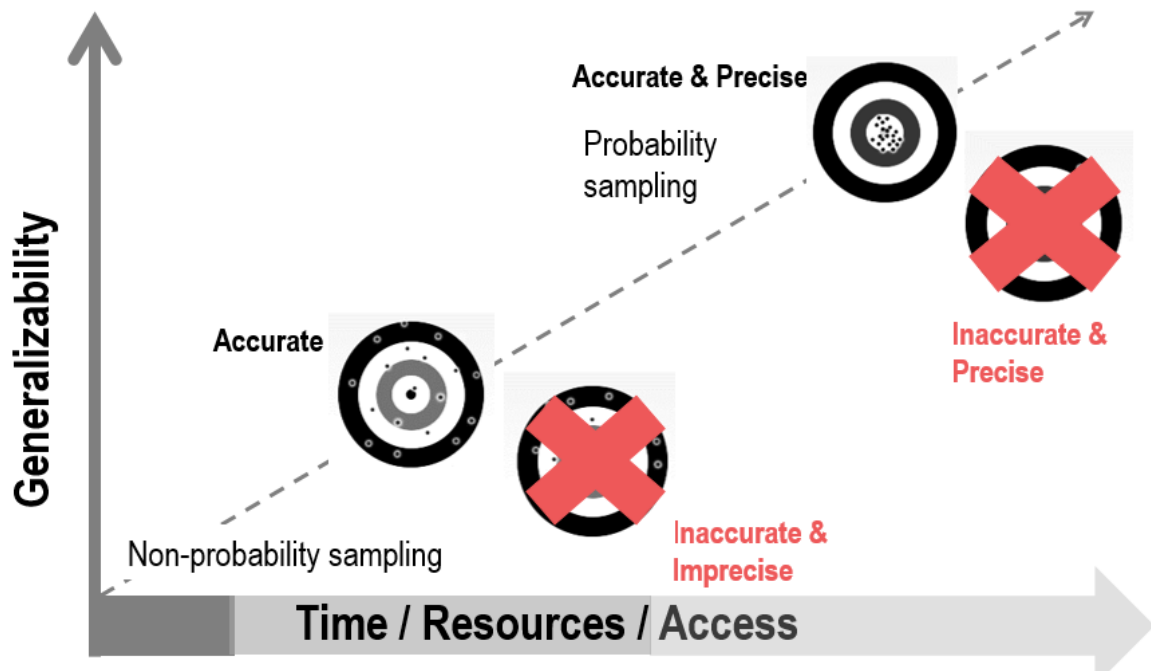
Ultimately, the level of **generalizability of research findings** will be **based on the type of sampling strategy** that is used (see Figure 1):

- a. **Probability sampling**
 - If done right i.e. randomisation followed through – both accurate and statistically precise
 - If done wrong i.e. selection biases introduced –statistically precise but inaccurate
 - False sense of preciseness is the worst as it generates misguided trust in findings → *should be avoided at all costs!*
- b. **Non-probability sampling**
 - If done right i.e. appropriate targeting and selection of respondents – accurate but imprecise
 - If done wrong i.e. inappropriate targeting and selection of respondents – inaccurate and imprecise

¹³ There are analytical techniques that can be used for post-stratification weighting (e.g. calibration) as a way to maximise representatives of a non-probability sample. If this is something that you would like to incorporate into your research design, please reach out to HQ Research Design and Data team to discuss.

¹⁴ Except in the case of respondent-driven sampling. See Table 2 for details.

Figure 1: Generalisability based on sampling type



2. Key parameters for sampling

2.1 Define geographical area and population of interest

- a. Geographical areas are commonly identified based on:
 - **Secondary data review** and known information gaps
 - **Information needs** of relevant stakeholders in country (affected areas in protracted crises or large-scale sudden onset emergency, in the interest of monitoring interventions, etc.)
- b. Population of interest may include everyone within the target geographical area or can be limited to a specific population group within it, for e.g. refugees in host communities – this **depends on the information needs**
 - For the ease of the sampling process, it is imperative to define clearly who exactly the population of interest will be from the outset.

Table 1: Checklist for the selection of geographical areas during research design¹⁵

When selecting an assessment location, make sure you take these priorities into account
<ul style="list-style-type: none"> <input checked="" type="checkbox"/> Area with greatest need <ul style="list-style-type: none"> ○ What areas have been reported as the worst affected or to have the greatest need? ○ What areas are normally the most vulnerable? <input checked="" type="checkbox"/> Area where research can have the greatest impact <ul style="list-style-type: none"> ○ Where does IMPACT or a partner organization already have capacity, including pre-established presence, partners, infrastructure and capacity at a global level? ○ Where is there a need for better coordination and information? <input checked="" type="checkbox"/> Area with current lack of information <ul style="list-style-type: none"> ○ Where are agencies assessing or responding? ○ What areas are being neglected?

2.2 Define unit of measurement

The unit of measurement is **the unit that will be used to record, measure and analyse observations/information collected** as part of the research effort. Units of measurement can be individual, family, household, location, community, facility, institution, etc. It is necessary to **define this unit from the outset**.

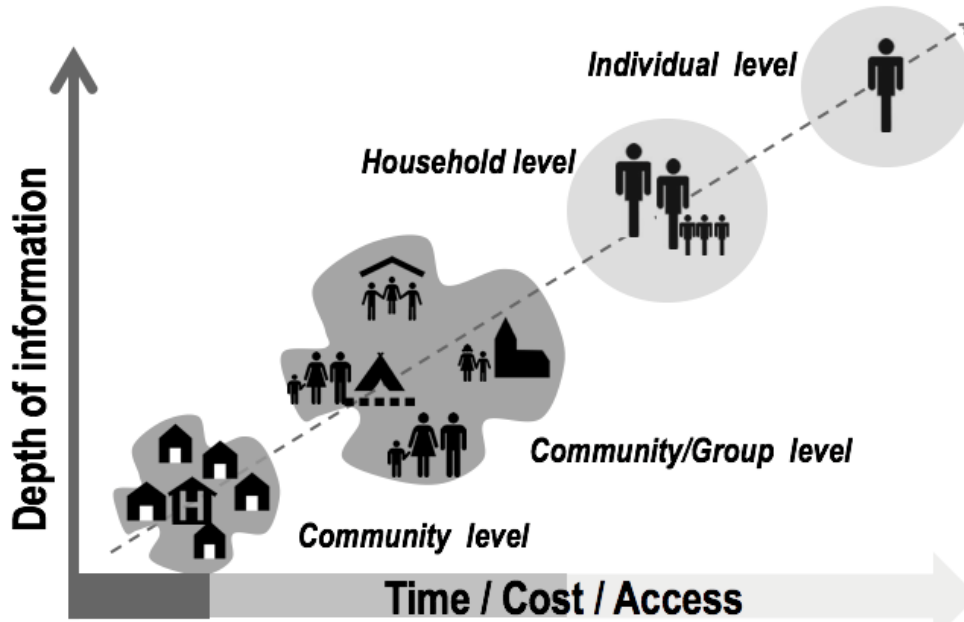
There are a few important things to keep in mind when defining this unit of measurement:

- Unit of measurement **impacts the time and amount of resources needed to collect and analyse information** → smaller the unit of measurement, smaller the volume.
- The unit of measurement selected is what will ultimately **define the depth and scope of the analysis** (see Figure 2). For example:
 - i. Individual level: What food items have *you* consumed over the past seven days?
 - ii. Household level: What food items have *your household* consumed over the past seven days?
 - iii. Community level: What food items have *the majority of people in this village* consumed over the past seven days?
 - iv. Institution level: What food items are currently available in *this market*?
- It is important that **different units of measurement are not conflated in the same data collection method**, for example, household survey questions for a village-level key informant questionnaire.

¹⁵ Adapted from CARE Emergency Toolkit (May 2009)

- For **rapid assessments**, information is typically collected at the **community or location level** to get the 'big picture' overview, before going in to more depth at group, household or individual level on identified issues.

Figure 2: Depth of information based on unit of measurement



3. Types and applicability of different probability / non-probability sampling strategies

As mentioned above, there are two categories of sampling strategies: probability and non-probability sampling. Both of these, in turn, have their own types of sampling strategies (see Figure 3). Detailed descriptions of these different strategies, as well as the applicability of each one, are discussed in detail in Table 2 below.

Figure 3: Overview of the types of probability and non-probability sampling strategies

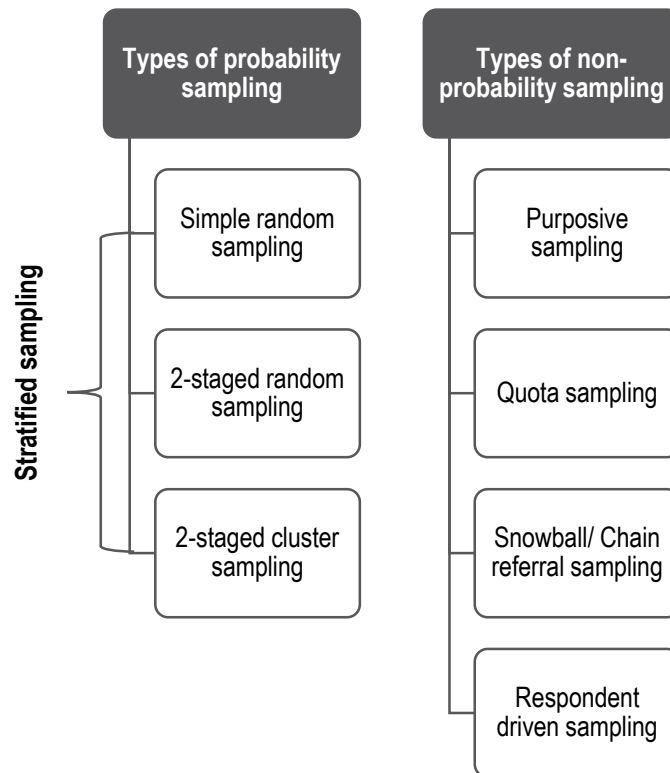


Table 2: Types and applicability of different sampling strategies

Category	Type of sampling strategy	Description	Advantages/ When to use	Disadvantages / Reasons to avoid
Probability	Simple (non-stratified) random sampling	A type of sampling when all units in the population of interest have an equal probability of being selected for the research. Typically, a complete list of all units are available to sample from.	Findings can be generalized to the entire population of interest with a known level of statistical precision.	Findings often will only show the average experience within the population of interest, when in reality there are quite a few variations to be taken into consideration.
Probability	Stratified random sampling	Similar to simple random sampling but involves stratifying the population of interest based on shared characteristics. Stratification means that specific characteristics of the population of interest (for e.g. demographic factors, geographical locations, socio-economic status, etc.) needs to be represented in the sample ; ¹⁶ to enable generalisation not only to the overall population, but also to each strata. For example: → Within the overall population of 1000 Syrian refugee households (HHs) in Jordan, we are also interested in understanding the specific situation of (1) HHs that arrived within the past one year vs. those that arrived more than one year ago, and (2) HHs residing in the north (where all the formal camps are) vs. those in the south (where refugees tend to reside mostly in informal tented settlements). → Stratified random sampling would mean we are dividing up the population of interest into four stratas – HHs in the north that arrived in the past year, HHs in the north that arrived > 1 year ago, HHs in the south that arrived recently, HHs in the south that arrived > 1 year ago – and then drawing a random sample within each of these stratum individually.	<ol style="list-style-type: none"> 1. Findings can be generalized to the entire target population of interest as well as to different subsets (strata) within the population, with a known level of statistical precision. 2. This also enables making comparisons between different groups and/ or geographical locations, as needed per the objectives of the research. 	<ol style="list-style-type: none"> 1. A larger sample size required compared to simple random, depending on the number of strata included. The more strata introduced, the larger the sample size will be. 2. If accurate population data and/ or location information is not available for one or more strata (see section 4 'Operationalising the selected sampling strategy' below), this sampling method often becomes quite challenging to implement.

¹⁶ Creswell, John W.; 'Research Design: Qualitative, Quantitative and Mixed Methods Approaches' (Third Edition, 2009); p.148

Category	Type of sampling strategy	Description	Advantages/ When to use	Disadvantages / Reasons to avoid
Probability	Stratified random sampling (continued)	Stratified sampling should always be used for an experimental approach to data collection . ¹⁷ At minimum, this involves stratifying the sample between a treatment group (i.e. a group that has received or will receive a specific treatment or intervention) and a control group (i.e. a group that has not received or will not receive the same treatment or intervention). The key purpose behind this is to evaluate impact of said intervention; by comparing the treatment and control groups the researcher can isolate whether it is the treatment and no other factors that have influenced a certain outcome. ¹⁸	See above	See above
Probability	2-stage random sampling (can be stratified or non-stratified)	Similar to simple random sampling but when a complete list of sampling units is not available for the area of interest (for e.g. beneficiary household lists or shelter footprint maps), population size per location is used to determine how many of the total surveys should be conducted in each location . As such, a household/ individual in an area with a bigger population has a higher chance of being selected than a household/ individual in an area with a smaller population. The location would typically be a smaller area/ administrative division within the wider area of interest (e.g. districts within a state). Additionally, depending on the population distribution within this wider area , locations which represent a very small proportion of the total population might not be assessed at all.	<ol style="list-style-type: none"> 1. Findings can be generalized to the entire population of interest with a known level of statistical precision. 2. Eases logistical planning by indicating which locations to target and how many surveys to conduct in each location, especially when the geographical area of interest is very widespread and the population distribution is relatively uneven. 	<ol style="list-style-type: none"> 1. Can result in a large number of locations to be visited, with some of them requiring a very small number of surveys (for e.g. <5 surveys) 2. If population data included in the initial sampling frame for locations is inaccurate, could create challenges during operationalization wherein data collection teams arrive at a location to collect x number of surveys, when the population of interest actually does not exist in that location or is of a much smaller size than was expected.

¹⁷ This is a research approach where independent variable(s) are manipulated and applied to dependent variables to measure their impact on the latter. Since one of the primary purposes of such an experimental research design is to detect an effect, it is important to factor in statistical power into the research design.

¹⁸ Creswell, John W.: 'Research Design: Qualitative, Quantitative and Mixed Methods Approaches' (Third Edition, 2009): p 146

Category	Type of sampling strategy	Description	Advantages/ When to use	Disadvantages / Reasons to avoid
Probability	<p>2-staged cluster sampling (can be stratified or non-stratified)*</p> <p>*See also: <i>IMPACT Cluster Sampling memo available on the online document Repository (Toolkit) here</i></p>	<p>Similar to random sampling except involves two stages: (1) first a primary sampling unit (PSU) is randomly selected¹⁹ with replacement, with the selection based on probability proportional to size (PPS)²⁰ i.e. probability of selection inverse to the population size of PSU and (2) the secondary sampling units (for e.g. households or individuals) are then selected within the randomly sampled PSUs. The number of units to be targeted in each PSU (i.e. number of households or individuals to survey) would be determined by the number of times the PSU is picked during first stage sampling.</p> <p>For example, if the population of interest is 15,000 Syrian refugee households nationwide:</p> <ol style="list-style-type: none"> <u>Stage 1</u>: Randomly select districts in (1) northern (2) southern governorates (region= strata, district= PSU) <u>Stage 2</u>: Randomly select Syrian refugee households within each PSU, with more surveys in districts with more refugees. <p>A key parameter for drawing a cluster sample is cluster size which pre-defines the minimum number of surveys to be done per PSU (5/10/15/etc.). As such, a PSU with a population size < predefined cluster size would not have a chance of being selected for the research.</p>	<p>Makes the scope of data collection more logistically feasible by reducing the number of locations where surveys need to be conducted. This is especially advantageous when the population of interest is very widely scattered across a geographical area.</p>	<p>2-staged cluster sampling suffers from 'design effect',²¹ which increases the number of units that need to be sampled to achieve the same level of precision as a random sample.²² The reason for this is that due to the shared environment within a cluster, units in a cluster tend to be more similar than units randomly selected across an entire population. For example, refugees in the same camp tend to face similar challenges in accessing livelihoods or have a similar view on which services are most in need of improvement. This means we obtain less information about the entire population from a given number of units in one cluster compared to the same number of units from the entire population. As such, the number of sampled units need to be increased to mitigate this lack of variance and obtain the same level of precision as a random sample.</p>

¹⁹ The PSU would typically be a smaller geographical area or administrative division within the wider targeted area.

²⁰ Probability proportional to size (PPS) is a method within sampling from a finite population in which a size measure is available for each population unit before sampling and where the probability of selecting a unit is proportional to its size. See also: Skinner, Chris J.; "Probability Proportion to Size (PPS) Sampling"; [Wiley Stats Ref: Statistics Reference Online](#) (August 2016).

²¹ Design effect is 'a coefficient which reflects how sampling design affects the computation of significance levels compared to simple random sampling'. See also: World Health Organisation (WHO), '[Steps in applying Probability Proportional to Size \(PPS\) and calculating Basic Probability Weights](#)'.

²² For cluster sampling, the target sample size from random sampling is adjusted for design effect, as outlined by Kish in 1965. This adjustment is made by applying the following formula: $n = n_{\text{eff}} (1 + (M - 1) ICC)$; where n_{eff} = effective sample size, $n_{\text{unadjusted}}$ = unadjusted sample size, M = average sample size per cluster, ICC = intra-cluster correlation.

Category	Type of sampling strategy	Description	Advantages/ When to use	Disadvantages / Reasons to avoid
Non-probability	Purposive sampling	<p>A type of sampling strategy when research participants and sites/ locations are purposefully selected based on what the researcher considers to be most appropriate to answer research questions.²³ Purposive sampling can be stratified or non-stratified.</p> <p>Some common types of purposive sampling:²⁴</p> <ul style="list-style-type: none"> → <u>Maximum variation/ heterogeneous sampling</u> i.e. a technique used to capture a wide range of perspectives, from the typical to the more extreme conditions; → <u>Homogenous sampling</u> i.e. a technique that aims to achieve a homogeneous sample; in other words, a sample whose units share the same or very similar characteristics (e.g. similar in terms of age, gender, occupation, etc.). This is useful to understand characteristics or conditions specific to a particular group of interest (for e.g. women of a certain age); → <u>Extreme / deviant case sampling</u> i.e. a technique that focuses on cases that are special or unusual, typically to highlight notable outcomes. This is useful when limited time, access and resources make it difficult to visit every single location and to reduce scope of data collection; → <u>Expert sampling</u> i.e. a technique that is used when the research needs to leverage knowledge from individuals that have particular expertise in some areas, typically through KI interviews (for e.g. WASH, agricultural practices, etc.). 	<ol style="list-style-type: none"> 1. If selection of participants is accurately done, information collected can be considered somewhat representative of the wider population of interest, although not with a known level of statistical precision. 2. Most appropriate type of sampling if we are collecting community level data. For example, if we want to know the total number of teachers in a school, we would ask a key informant (the head-master) rather than drawing a random sample of teachers to estimate this. 3. Can be a suitable non-probability alternative if probability sampling not logistically feasible. 	<p>Prone to researcher biases, which limits the ability to make generalisations to the wider population of interest. In other words, results can be considered indicative but not statistically representative.</p> <p>Despite this, depending on the research objectives, there are instances where purposive sampling is relevant simply because statistical representativeness is irrelevant. For example, if we are conducting a KI interview with a camp WASH technician because we want to know what the main issues are with the water network, then purposive sampling is more relevant than non-probability sampling (we would not get better precision by randomly selecting someone to talk about the water network).</p>

²³ Creswell, John W.; 'Research Design: Qualitative, Quantitative and Mixed Methods Approaches' (Third Edition, 2009); p.178

²⁴ See also: <http://dissertation.laerd.com/purposive-sampling.php>

Category	Type of sampling strategy	Description	Advantages/ When to use	Disadvantages / Reasons to avoid
Non-probability	Quota sampling ²⁵	Non-probability version of stratified sampling where a target number of interviews - a quota - is determined for a specific set of homogenous units (for example, based on gender, age, location, etc.), with the aim of sampling until the respective quotas are met. The quotas should be set to reflect the known proportions within the population . For example, if the population consists of 35% female and 65% male, the number of FGDs or interviews conducted with males and females should also reflect those percentages.	<ol style="list-style-type: none"> 1. If selection of participants is accurately done, information collected can be considered somewhat representative of the population groups of interest, although not with a known level of statistical precision. 2. Can be a suitable non-probability alternative if stratified random sampling or stratified cluster sampling is not logistically feasible. 	Same as above.
Non-probability	Snowball sampling	<p>A sampling strategy wherein households or individuals are selected according to recommendations from other informants and research participants. Each participant recommends the next set of participants to be contacted for the study.²⁶ Snowball sampling can be both stratified or non-stratified, depending on the research needs.</p> <p>Snowball sampling is also sometimes referred to as 'chain referral' sampling.</p>	<ol style="list-style-type: none"> 1. Can be a suitable alternative to purposive sampling when the population of interest is hard to reach from the outset and/ or could be hesitant to participate in the assessment. 2. Could be cheaper and less time-intensive in terms of planning, in comparison to probability sampling or other non-probability sampling strategies. 3. Can be a good means of implementing purposive sampling, for e.g. asking KIs we interview to help put us in touch with KIs from specific areas and/ or with specific knowledge 	<ol style="list-style-type: none"> 1. Researcher has limited control over the final sample. 2. Prone to respondent biases, which limits the ability to make generalisations to the wider population of interest. In other words, results can be considered indicative but not statistically representative. 3. Respondent biases can also lead to a significant over-representation or under-representation of a specific group within the wider population of interest, thus skewing the results.
Non-probability	Respondent-driven sampling (RDS) ²⁷	A variation of snowball/ chain referral sampling which uses social network theory ²⁸ to overcome the respondent bias limitations associated with snowball	<ol style="list-style-type: none"> 1. Retains advantages of non-probability sampling, especially snowball sampling, while mitigating 	<ol style="list-style-type: none"> 1. Requires a relatively larger sample size than other non-probability sampling strategies.

²⁵ See also: Brown et al; '[GSR Quota Sampling Guidance](#): What to consider when choosing between quota samples and probability-based designs' (UK Statistics Authority, 2017)

²⁶ IMPACT Initiatives; '[Area-based Assessment with Key Informants: A Practical Guide](#)' (December 2018); p.6

²⁷ For a detailed understanding of respondent-driven sampling, see also: WHO & UNAIDS; '[Introduction to HIV/ AIDS and sexually transmitted infection surveillance Module 4: Introduction to respondent-driven sampling](#)' (2013).

²⁸ Social network theory is a branch within sociological statistics which aims to map relationships and characteristics shared by groups within a population of interest

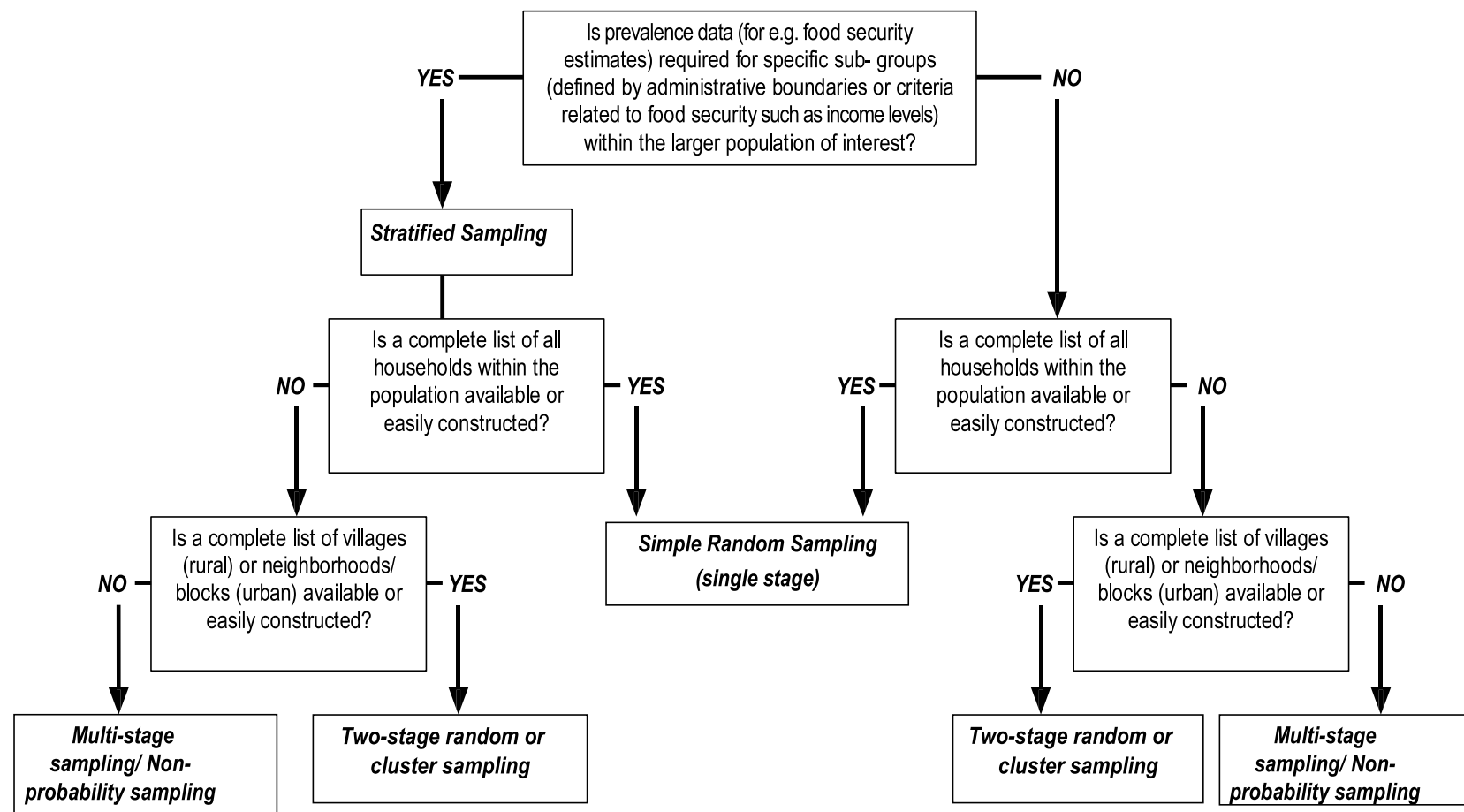
		<p>sampling. Specifically, RDS uses information about the social networks of participants recruited to determine the probability of each participant’s selection and mitigate the biases associated with under sampling or over sampling specific groups.²⁹</p> <p>RDS comprises of different steps:</p> <ul style="list-style-type: none"> → <u>Initial recruitment</u>: an initial identification and recruitment of participants who serve as the ‘seeds’ from the population of interest. A diverse selection of seeds from the outset will help ensure reaching diverse members of population by the end. → <u>Recruitment chain follow-up</u>: starting from the ‘seeds’ generate long recruitment chains made of several recruitment waves of participants so that the final sample characteristics will be independent of those selected as ‘seeds’. → <u>Analysis component</u>: a careful collection of personal network size information and tracking who recruited whom is critical for the analysis of RDS data.³⁰ <p>Contrary to other non-probability sampling strategies, calculating RDS sample size needs to factor in the following: design effect, estimated proportion (to test prevalence at one time), desired level of change in the measures of interest (over time), level of significance and level of power desired.³¹</p>	<p>limitations associated with respondent bias.</p> <p>2. Increases ability to have representative findings, generalizable to the population of interest with a specified level of precision, in comparison to other non-probability sampling strategies.</p>	<p>2. Functionality of RDS is based on some key assumptions, mainly that respondents within the population of interest know one another and are linked by some sort of component. As such, additional time needs to be factored in at planning stage for a formative assessment to (1) explore social networks within the population of interest to determine whether peer-to-peer recruitment can be sustained by the survey population and (2) identify the ‘seeds’ for the initial recruitment phase.</p>
--	--	--	---	--

²⁹ WHO & UNAIDS; 'Introduction to HIV/ AIDS and sexually transmitted infection surveillance Module 4 Unit 1: Introduction to respondent-driven sampling' (2013), p.17-25

³⁰ For more detailed understanding of social network analysis, see also: IMPACT Initiatives; 'Area-based Assessment with Key Informants: A Practical Guide' (December 2018).

³¹ For a detailed understanding of how to calculate sample sizes for respondent-driven sampling, see also: WHO & UNAIDS; 'Introduction to HIV/ AIDS and sexually transmitted infection surveillance Module 4 Unit 3: Sample size calculation for RDS' (2013), p.45-54

Figure 4: Decision tree for choosing an appropriate probability sampling strategy^{32,33}



³² Adapted from World Food Programme (WFP), 'Sampling Guidelines For Vulnerability Analysis' (2004)

³³ For stratified sampling, when we say "a complete list" is available, this list does not necessarily have to be in a list form but just has to contain all units in the population of interest (for e.g. a map of villages or households in a camp setting).

4. Operationalising the selected sampling strategy

Once the appropriate sampling strategy has been selected and agreed upon, the following steps need to be taken to finalise the sampling and overall methodology:

- 1) Prepare sampling frame
- 2) Calculate sample size
- 3) Finalise strategy to select units within sampling frame i.e. identify participants for data collection

4.1 Prepare sampling frame

Sampling frame is essentially **a list of all units in the population of interest that is used to draw the sample.**

This can either be (1) **a complete roster** of all individuals or households (depending on the unit of measurement) within the area of interest (2) a list of the **size of the population of interest** or (3) **a map** of communities or households within a camp or village. For instance, if our population of interest is Syrian refugee households in Jordan, stratified by region (with refugees in the north living in both in formal camps and outside) and time of arrival, the sampling frame could be as shown in Table 3 below.

Table 3: Example sampling frame for refugee households in Jordan, stratified by region and time of arrival

	Households arrived within the last 1 year	Households arrived > 1 year ago
North Jordan (in camp)	550	3,600
North Jordan (out of camp)	1,500	6,350
Central Jordan	850	2,200
South Jordan	980	3,500

4.2 Calculate sample size

Once the sampling frame has been defined, how do we use this to calculate the required sample size? This will differ based on the type of sampling strategy i.e. probability or non-probability.

- The **required size of a probability sample** (e.g. number of household or individual surveys to be conducted) is calculated **based on probability theory** and on the **target level of statistical precision required** for the research findings (see box below).

Calculating the sample size for a probability sample

A lot of online tools already exist to help with this calculation. **IMPACT's own in-house sampling tool to calculate sample sizes is available through [this link](#).** Please see Annex 2: User Guide for IMPACT's online sampling tool.

You want to conduct a survey, with a representative sample of households in a refugee camp of 550 households.

- How many households should you survey to ensure that your findings have a confidence level of x% and an error margin of +/-y%?
- The formula used within IMPACT was first outlined by Krejcie and Morgan in 1970. The formula is $n = \chi^2 N p(1 - p) / \beta^2 (N - 1) + (\chi^2 p (1 - p))$; where n=sample size, χ^2 = Chi-square for the specified confidence level at 1 degree of freedom, N=Population size, P= Population proportion (assumed to be 0.5 to generate maximum sample size), β = desired Margin of Error (expressed as proportion)
- For an experimental survey design, statistical power may also need to be factored in for sample size calculation.
- If population size is unknown, an infinite population can be assumed to draw the required sample size. The risk with this is oversampling and giving unequally weighted representation to an unequally distributed population.

- Contrary to probability sampling, **sample sizes for non-probability sampling are calculated based on what is feasible and what should be the minimum to meet the research objectives** with quality standards.³⁴

This can be done in the following ways:

- **Sample size based on feasibility** i.e. the maximum possible given time, access and resources available.
- **Sample size led by saturation** i.e. continue conducting interviews and discussions until data saturation has been achieved and no new themes or discussion points are appearing in the data that is being collected. See saturation grid example in Table 4 below.
- **Sample size based on what is known of the population of interest** i.e. setting targets based on specific characteristics such as population size and demographic breakdown. For instance, if we want to conduct FGDs to understand a population's ability to access basic services across three different districts, it would make sense to: (1) conduct from the outset a minimum of two FGDs per district, one male and one female; and (2) conduct two additional FGDs in District 2 because it also has a large internally displaced population whose experiences may be different from the overall non-displaced population.

Table 4: Example saturation grid for data collection using non-probability sampling

FGD ID	1	2	3	4
# FGD participants	5	6	8	7
Main PULL Factors_ Discussion Point (DP)1: [Security]	y			y
Main PULL Factors_ DP2 [Food availability]				y
Main PULL Factors_ DP3 [Water availability]		y	y	y
Main PUSH Factors_ Discussion Point (DP)1: [Insecurity]	y			y
Main PUSH Factors_ DP2 [Lack of food]				y
Main PUSH Factors_ DP3 [Lack of water]		y	y	y
Etc.				
Etc.				
Total # of new DPs added	2	2	0	2

4.3 Finalise strategy to select units within sampling frame i.e. identify participants for data collection

This will need to **take into consideration all available information of the population of interest** such as where they are located on the ground and how they can be reached.³⁵

The strategy used for finding research participants **varies for probability and non-probability sampling**.

³⁴ Except in the case of respondent-driven sampling. See Table 2 for details.

³⁵ Creswell, John W.; 'Research Design: Qualitative, Quantitative and Mixed Methods Approaches' (Third Edition, 2009); p.148

- Probability sampling: The key here is **randomization** i.e. ensuring equal chance for all units to be selected for the research. **Any bias introduced in the selection process** compromises the extent to which the findings can be **considered truly representative** of the population of interest.

Table 5: Strategies for random selection of research participants – probability sampling

	Description	Pre-requisites
[Option 1] List-based random selection	Random units are selected from a list (including mapped shelter points) containing the entire population of interest.	An accurate, up-to-date list of all units in your population of interest with required details is easily available (for e.g. population list with location points, beneficiary list with contact details, etc.)
[Option 2] Random selection on site- GIS sampling	Random GPS points are generated on a map covering the population of interest. The distribution of GPS points is weighted based on population density, should this vary across the targeted area. A unit located nearest to each point (within a pre-defined buffer as relevant to context) is then targeted for the survey. See Annex 3 for a detailed note developed by the REACH Jordan team in May 2016 with guidance on how to implement GIS-based sampling.	<ol style="list-style-type: none"> 1. Accurate, up-to-date shape files for administrative boundaries are easily available 2. Reliable data indicating the distribution of the population and population density across the targeted area is easily available 3. Well-trained data collection teams that have the capacity to use maps.me or similar navigation software to locate sampled GPS points on the ground.
[Option 3] Random selection on site- Systematic sampling	Systematic measures are taken on site to ensure that the entire radius of the targeted area is covered and all units within this area have a probability of being selected. See FAQs section below (page 24) for two examples of systematic measures that have been used by IMPACT teams across different contexts.	<ol style="list-style-type: none"> 1. Accurate understanding of the layout of the area to be targeted (for e.g. boundaries of sites/settlements) 2. Area is of a manageable size to implement systematic sampling; otherwise, it will need to be broken down into sub-areas (for e.g. camp blocks or city neighbourhoods) to implement systematic sampling

- Non-probability sampling: Unlike probability sampling, there are **no structured or systematic rules or methods** governing the selection of research participants for non-probability sampling. However, the **following key things should be considered** in the selection of participants:
 - ➔ If you can **select just one** participant, who should you select?
 - ▶ Who would be the **most 'representative' or most 'typical'** participant to provide a good understanding of the topic of investigation?
 - ➔ How many participants do you need to ensure the information collected is **as accurate as possible**?
 - ▶ Are there any additional participants you should consider to **ensure the disadvantaged or minorities** within the wider population of interest are also well-represented in the sample?
 - ▶ How can you ensure **some variation in the perspectives and views represented** even if the sample is somewhat homogenous? For example, ensuring representation of different age groups in a focus group of female refugees.
 - ▶ Can one type of informant or participant give you all the information you need or should you **target different types of profiles** to fill out the same questionnaire?
 - ➔ Should some profiles of participants be given **more weight than others** due to their level of knowledge and/ or ability to provide information on a specific topic?
 - ▶ This is especially useful **when participants contradict each other's responses** for the same unit of measurement (for e.g. two different KIs providing contradictory information on the same camp). The pre-defined weights help triangulation of responses in these instances.

5. Frequently asked Questions (FAQs)

5.1 FAQs on choice of sampling strategies

a. When should I use probability sampling over non-probability sampling?

→ This decision should be based on the following key considerations:

- i. **Research questions and objectives** i.e. do these require identifying and measuring prevalence of attributes of a wider population and making generalizable claims of this? *If yes → probability sampling is better.*
- ii. **Economy of design** i.e. do you have the required level of access and the necessary resources (both human and material) to implement probability sampling (i.e. potentially access any of the areas where your population of interest is present)? *If no → probability sampling may not be possible to implement in a robust way.*
- iii. **Time available** i.e. can the required scope of data collection and analysis be completed with the time and capacity available? Probability sampling usually takes longer to implement. *If no → probability sampling may not be possible to implement in a robust way.*
- iv. **Robust sampling frame** i.e. is accurate information available, or can be collected (location, population size, etc.) for the population of interest? *If no → probability sampling will not be possible to implement robustly.*

b. When should I use 2-staged cluster sampling over random sampling?

→ 2-staged cluster sampling is often more beneficial to be used when it is logistically difficult to access the population of interest (for example, because this population is too widely scattered across the geographical area to be assessed). The final decision should be based on a simple cost-benefit analysis: estimate the target 'ideal' level of precision (for e.g. 95/5) to identify your effective sample (e.g. 385) for random sampling, then calculate whether it would demand more resources / time to (1) visit lesser locations but conduct more surveys (2-staged cluster sampling) or (2) visit more locations but conduct less surveys (random sampling).

c. I need to have representative data by geographical location and/ or population group but there are too many administrative units in my context (for e.g. 500+ districts) which is substantially increasing my sample size upon introducing this stratification. What should I do?

→ *[If we only need stratification by geographical location]* You could consider the **higher administrative level** than what you were initially considering (for e.g. governorates instead of districts). However, sometimes going just by the higher administrative level may not work as the findings are still required at the more granular administrative level (i.e. district). In this instance, the solution is to **group up districts by certain shared characteristics** which may or may not be directly related to the geographical distribution and proximity of the districts. These groups rather than the individual districts can then serve as the strata for sampling purposes, based on the assumption that all districts within a group will have similar experiences. Some of the examples of the types of characteristics that have been used previously for such groupings include: population size, socio-economic characteristics, livelihood zones, and agricultural production trends. The key here is to ensure the characteristics used for grouping are carefully considered and discussed so that they do indeed reflect the situation and shared experiences vis-à-vis the topic being studied on the ground.

→ *[If we need stratification by both geographical location and population group]* The solution here would be to **draw an un-stratified sample per geographical level** (for example, all population groups combined at Admin 3) **AND stratification by population group at the level higher than that** (for example, findings for both IDPs and non-IDPs at Admin 2). If respondents at Admin 3 level are truly randomly selected, when the sample is aggregated, required sample sizes could be achieved per

population group at Admin 2 level. If population data by group is available at Admin 3 level, it is also possible to factor this into the Admin 3 sample i.e. ensure that the overall sample for each Admin 3 is distributed based on the population group breakdown within it. Finally, if there is reason to assume that these Admin 2 population group sampling targets cannot be met through a simple aggregation of Admin 3 samples (for e.g. because a specific population of interest is difficult to find / concentrated in very specific areas), a top-up sample can be drawn by population group from the outset, which can be used to collect the remaining sample at Admin 2 level.

d. Can I still claim to have representative findings with non-probability sampling?

- Due to inherent biases and room for error with sampling design, findings from a non-probability sample cannot be considered representative with a known level of statistical precision. However, **if the research locations and research participants are carefully selected**, findings can still be considered at least somewhat representative, even if we can't calculate the exact level of precision. For location selection, some of the key things to keep in mind to have wider representation is:
 - i. Ensure wide coverage i.e. collect information from as many locations and communities as possible within a district or sub-district rather than from one or two locations only
 - ii. Ensure variation between locations from which information is collected; i.e. collect information from different types of locations to have wider representation of different groups and experiences

5.2 FAQs on operationalising sampling strategies

e. What additional points do I need to keep in mind for sampling if I am conducting my data collection via phone rather than face to face?

- See Annex 4: Sampling considerations for remote phone-based data collection

f. When drawing a cluster sample, for some strata, the number of surveys I am meant to conduct is significantly higher than the random sample. Why is this case? How can I address this?

- The higher number of surveys is understandable because of the **design effect associated with 2-staged cluster sampling**. In some cases, the design effect can be exacerbated, when within a stratum, there are specific PSUs that have significantly higher population sizes (for e.g. densely populated urban centers) than others. These PSUs will have a high number of surveys to be conducted in a cluster, which inflates the design and by consequence the overall sample size for the strata. The solution to overcome this would be to separate out these PSUs into a distinct stratum and draw the sample for this stratum separately from the other PSUs. The sample from the distinct strata and the remaining PSUs can then be aggregated to have the sample for the overall strata.

g. For an experimental survey design, what are the key things I need to keep in mind when defining the sampling frame for my control group?

- The most important thing to factor in is to identify the control group **based on certain shared traits or characteristics with the treatment group**, which would make this control group comparable with the treatment group. For example, if the treatment group comprises of host communities in a specific region of the country, the control group could be host communities in the same region who are not receiving the same treatment. Alternatively, the control group could also be host communities in another region that are (1) also hosting refugees and (2) had similar socio-economic conditions and service provision capacities prior to the arrival of large numbers of refugees in the country.

h. What if I don't think the population data available from secondary sources for my sampling frame is reliable, accurate and up-to-date?

→ Before drawing and implementing the sample, it is imperative to ensure that the population data being used is reliable, accurate and up-to-date. Otherwise, this would complicate the ability to implement sampling on the ground and also turn into a logistical nightmare if data collection teams are unable to locate the population of interest in targeted areas. A key good practice to overcome this is to conduct a **preliminary scoping exercise** during the planning phase to verify the sampling frame; for e.g. conduct some KI interviews to clarify that the population of interest is indeed distributed in the way the secondary data sources are saying it is. Alternatively, the data should be **triangulated with other sources**, including population density data available from satellite imagery.

i. What do I do if the randomly sampled GPS point falls at a point where there are no households to survey or there is no eligible respondent at the time of data collection?

→ A **pre-defined radius** (for e.g. within 10 meters around sampled point) can be set within which the enumerator is able to locate a household to survey in these instances. If the radius also does not work, **a buffer of GPS points should always be available from the outset** to ensure the required sampling targets can still be met. It is important that randomization is retained in these instances, the same rule for randomization is followed throughout, and some kind of snowballing is not used to identify the "eligible" household as this would compromise the probability of the sample.

j. What do I do if the randomly sampled GPS point falls at a point where there is a multistoried building and multiple households living in it?

→ It is very important that a **clear rule for this is established** with the data collection teams before data collection begins. For example, a random number generator can be used at the GPS point to determine (1) which floor and (2) which apartment/ household on this floor should be interviewed.

k. What are some methods I can use for systematic sampling on site?

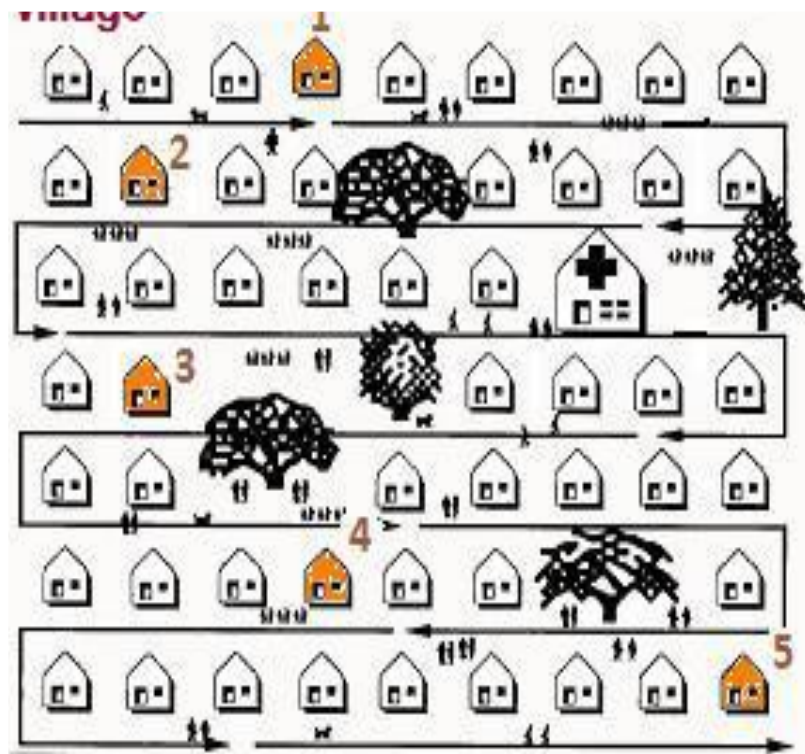
→ This depends on the type of location the sampling needs to be undertaken for. **A step-by-step guide on one method** that has been used by IMPACT teams across different contexts is outlined below:

- i. Enumerators meet at the center of the targeted location (village/ site/ settlement), spin a pen and each enumerator starts walking in a direction towards the edge of the location as shown by the pen
- ii. On his/ her way to the edge, he/ she counts either the number of households passed OR the time taken to reach the edge (depending on how big the location is, in a bigger location with many households it makes more sense to count the time)
- iii. Once he/ she reaches the edge they then determine the threshold for which household to interview on the route based on: # of HHs in the route or time taken to reach the edge / the target # of HHs to be interviewed per enumerator
- iv. The enumerator then starts walking back towards the center and assesses every xth household (with x as determined by the formula in point #iii above).

Another method that could work for locations that are organized in a clearly structured way (for e.g. camp with same number of households per row) is listed below (see Figure 5):

- i. Calculate a threshold based on total population in location (let's say 60 households) / sample needed from the location (let's say 5) → $60/5= 12$
- ii. From the starting point of the location, select the first household randomly between 1 and 5
- iii. After the first household, interview every 12th household following a single direction in a clearly laid out route until the edge of the settlement has been reached.

Figure 5: Systematic selection on site – Option 2



I. How do I identify a respondent in a random household survey?

- The typical approach for a household survey, especially in humanitarian contexts, is to gather information from the **head of household** or – in his/ her absence – any other adult household member that is knowledgeable about the affairs of the household. The definition of “head of household” should be very clear to data collection
- Sometimes, however, it might be important to **speak to someone other than the head of household for specific parts of the survey**, for e.g. a female member to discuss their protection concerns or to understand household’s female hygiene situation.
- In some contexts, like Bangladesh, IMPACT/ REACH teams have also in the past included an additional layer to ensure households are randomly selected in a way that is clear where a **female vs. male respondent from the household** should be interviewed. Please see box below for details from the Bangladesh example. This particular approach is useful if there is a hypothesis or assumption to be tested that information being collected through the survey tends to vary based on whether the respondent is male or female. For example, if we are conducting a protection assessment with a module on gender-based violence, in some contexts, issues at household level may be under-represented in the findings if we were only to speak to male heads of households.

Box 1: Selection of respondents according to gender

According to the 2017 Inter-Agency Standing Committee Gender Handbook for Humanitarian Action, in-depth joint assessments should “be aware of possible biases in information collection and analysis. For instance, if women [are] not consulted, the identified priorities do not reflect the needs and priorities of the whole community.” Traditionally, household survey methodology dictates that the—predominantly male—head of household is interviewed on behalf of the entire household. However, preliminary information needs for this assessment identified by ISCG sectors involve a mix of objectively verifiable household-level indicators (such as coping strategies or water points, etc.), and perception-based indicators that are likely to reflect the specific opinions of the individual respondent (such as sense of safety, perceptions of relationship with Rohingya refugees, etc.). In these circumstances, constraining respondents to male household heads is likely to significantly bias the sample in favour of men’s views.

This assessment will therefore adopt a “good enough” approach that aims to balance the need to obtain a sample that is randomised in order to be generalisable to the entire population and accurately reflects the situation of each household, with the need to obtain a gender-balanced overview of priorities and needs. Under this approach, each enumerator in a 50/50 gender-balanced team will ask to interview the single household member who is most knowledgeable about the affairs of the household, who is of the same gender as the enumerator. All interviews will be drawn from the same set of sample points (i.e. there will not be a separate set of “male” and “female” sample points). Under this approach, the sampling design is expected to yield data that can be compared according to gender of respondent at 95% margin of error and 10% confidence at the union level, as well as data that are generalisable to households across the Union at 95% confidence level and 6% margin of error.¹ A small number of indicators related to GBV will only be asked of female respondents, meaning that data for these indicators will only be generalisable to 95% confidence level and 10% margin of error at the union level. The compromises inherent in this approach with respect to randomisation will be clearly outlined in the limitations section of the study’s final report.

¹ An alternative approach of splitting household interviews by gender, with men and women answering different questions, was rejected as impracticable in terms of both the increased number of enumerators required, and the possibility of rejection of this approach by the target population.

m. What if I need to randomise respondent selection within the sampled household, for e.g. for an individual perception survey? How can I do this?

- If it’s two respondents to choose from, keep it simple and easy to implement- **a coin toss** will suffice.
- If it’s more than two respondents to select from, the **Kish grid approach** can be used. This is used in social science research to randomly select respondents within a household. For more detailed guidance on how to operationalize the Kish grid approach is available on [this link](#). Of course, the risk with this is (1) it requires training of enumerators on how to go about this and piloting to ensure training has been successful (2) it adds time to the survey and data collection process and (3) it is impossible to know for sure if enumerators are actually applying it all the time.
- If it’s more than two respondents to select from, **a random number selection module can be introduced in the Kobo form which follows a similar principle as the Kish grid approach**. To implement this approach: (1) all potential respondents within the household are asked to line up in a straight line (2) Kobo generates a random number (x) for you based on the total pool (3) you can then select the respondent standing xth in the line, where x is determined by the random number generator function on Kobo. (Note: There are past IMPACT assessments that have used this approach and therefore coded this function into the Kobo form. Please reach out to the IMPACT HQ Research Department if you want to do something similar and want access to this Kobo tool).

n. What is the ideal number of participants to have in an FGD?

- Based on IMPACT’s experiences over the years across different contexts, **a maximum of 6-8 participants** is the optimum for a constructive group discussion. Any smaller and the discussion does not yield the in-depth information desired. Any larger and the discussion becomes difficult to facilitate and it becomes challenging to gather the desired information with sufficient details.

o. How do I identify participants for an FGD?

- This can be either (1) **pre-arranged** (ideal option) whereby existing networks in the community of interest are leveraged to arrange the group of participants prior to the date of collection, based on clear pre-defined selection criteria specified by the research team or (2) done on site by field coordinators who **purposively select community members** on the day of data collection team based on pre-defined selection criteria aligned with the sampling strategy and research objectives.

5.3 FAQs on mitigating methodological issues encountered during data collection

p. I used 2-stage random sampling to calculate number of surveys needed per PSU. However, the population data used for sampling was found to be incorrect which means we may have conducted more surveys in some PSUs with a smaller population, or less surveys in some PSUs with a larger population than expected. What should I do?

- The main implication for this in terms of representation is that at the overall strata / area level, these PSUs will be inaccurately over-represented or under-represented, thus potentially skewing the results. It is important therefore that the updated population figures are used to apply weights at PSU level during the analysis, to mitigate this over or under representation.

q. What if I need to delete some data points or complete entries during data cleaning, which raises the risk of not meeting the required sample target?

- A sufficient buffer (10%, 15%, 20%, etc.) should always be included to mitigate this issue. This way, as long as the surveys “lost” falls within the buffer, you will still have the required sample size.

r. What if some sampled PSUs become inaccessible, for instance due to security reasons, during data collection?

- See Annex 5: Troubleshooting issues encountered due to inaccurate information in sampling frame

s. I have started processing the preliminary data coming in for a study that used a non-probability sampling approach and have realised that different respondents for the same unit (for example, KIs for a village) are providing contradictory information to the same question. What should I do?

- While it depends on the type of question, the first answer here is **triangulation**. This can either be triangulating with other available secondary data to see which of the responses are closer to what is already known, or triangulating based on pre-determined weights if a certain respondent profile was considered to be more ‘knowledgeable’ about the subject matter during the sampling design. Finally, as a last resort, an additional interview can be conducted to see which of the two contradictory responses are closer to this new data source.
- However, for some questions if there is **no clear consensus** established in the responses, rather than triangulating with weights, it might make sense to either not report on that variable OR go with an approach where for specific variables (e.g. safety and security issues in the village) we report the issues even if one of the KIs report it.
- In either case, a clear **aggregation and triangulation plan** should be in developed as part of the data analysis plan.

6. Annexes

Annex 1: Memo on cluster sampling

Available on the IMPACT Online Document Repository > Toolkit > Research > Research design > Guidance: https://www.impact-repository.org/wp-content/uploads/2018/11/Annex-4-Cluster_sampling-memo_modif_20200514.docx

Annex 2: User Guide for IMPACT's online sampling tool

1. Description

1.1. Functionalities

The probability-sampling tool can help with sampling for the most common strategies used in IMPACT research cycles. The tool is implemented in R and is hosted on a shiny server. Depending on the monthly usage, you may have to use one of the links below:

- https://impact-initiatives.shinyapps.io/r_sampling_tool_v2/
- https://oliviercecchi.shinyapps.io/R_sampling_tool_v2/

You can also run the tool offline with the code available on GitHub: <https://github.com/oliviercecchi/Probability-sampling-tool>.

The tool can generate sample based on fix number of survey or based on the sampling frame provided; the tool will calculate the sample size to reach the desired level of confidence and error margin on the findings and buffer as specified.

Different type of sampling can be performed:

- Random sampling from a list of households (can be stratified or not) – without replacement
 - o See section 4 of the sampling guidance, under “simple random sampling”
- Random sampling from a list of location and population by location – primary sampling unit (PSU) selected with replacement
 - o See section 4 of the sampling guidance, under “2 stage random sampling”
- Cluster sampling from a list of location and population by locations – PSU selected with replacement
 - o See section 4 of the sampling guidance, under “2 stage cluster sampling”

All the sampling types can be stratified by one factor (present in the sampling frame).

Example: In this sampling frame, a combination of district and population, group is used for stratification.

VDC name	District	VDC P_CODE	Population group	Stratification	Total number of HH
Bageswari	Sindhupalchok	C-BAG-26-001	Host	Sindhupalchok - Host	1137
Bageswari	Sindhupalchok	C-BAG-26-001	IDPs	Sindhupalchok - IDPs	2488
Balkot	Sindhupalchok	C-BAG-26-002	Host	Sindhupalchok - Host	3999
Balkot	Sindhupalchok	C-BAG-26-002	IDPs	Sindhupalchok - IDPs	2878
BhaktapurN.P.	Sindhupalchok	C-BAG-26-003	Host	Sindhupalchok - Host	17639
Changunarayan	Sindhupalchok	C-BAG-26-004	Host	Sindhupalchok - Host	1374
Chhaling	Sindhupalchok	C-BAG-26-005	Host	Sindhupalchok - Host	1817
Chitapol	Sindhupalchok	C-BAG-26-006	Host	Sindhupalchok - Host	1274
Dadhikot	Sindhupalchok	C-BAG-26-007	Host	Sindhupalchok - Host	2688
Duwakot	Sindhupalchok	C-BAG-26-008	Host	Sindhupalchok - Host	2412

Duwakot	Sindhupalchok	C-BAG-26-008	IDPs	Sindhupalchok - IDPs	2411
Gundu	Sindhupalchok	C-BAG-26-009	Host	Sindhupalchok - Host	1257
Gundu	Sindhupalchok	C-BAG-26-009	IDPs	Sindhupalchok - IDPs	15314
Jitpurphedi	Kathmandu	C-BAG-27-027	Host	Kathmandu - Host	1631
Daxinkali	Kathmandu	C-BAG-27-015	Host	Kathmandu - Host	29126
Dhapasi	Kathmandu	C-BAG-27-016	Host	Kathmandu - Host	4692
NaikapNayaBhanjyang	Kathmandu	C-BAG-27-042	Host	Kathmandu - Host	3919
Satungal	Kathmandu	C-BAG-27-049	Host	Kathmandu - Host	20302
Gonggababu	Kathmandu	C-BAG-27-022	Host	Kathmandu - Host	5027
Jitpurphedi	Kathmandu	C-BAG-27-027	IDPs	Kathmandu - IDPs	973
Daxinkali	Kathmandu	C-BAG-27-015	IDPs	Kathmandu - IDPs	3731
Dhapasi	Kathmandu	C-BAG-27-016	IDPs	Kathmandu - IDPs	1225
NaikapNayaBhanjyang	Kathmandu	C-BAG-27-042	IDPs	Kathmandu - IDPs	26395
Satungal	Kathmandu	C-BAG-27-049	IDPs	Kathmandu - IDPs	2278
Gonggababu	Kathmandu	C-BAG-27-022	IDPs	Kathmandu - IDPs	7817
TOTAL					163804

1.2. Sample size calculation

The formula used by IMPACT to calculate target sample sizes for probability sampling strategies within this tool was outlined by Krejcie and Morgan in 1970 and has been widely used in social research (3,313 known citations)³⁶ It is described as follows:

$$n = \chi^2 N p(1 - p) / \beta^2 (N - 1) + (\chi^2 p (1 - p))$$

Where:

n = Sample size

χ^2 = Chi-square for the specified confidence level at 1 degree of freedom

N = Population size

p = Population proportion (assumed to be 0.5 to generate maximum sample size)

β = desired Margin of Error (expressed as proportion)

1.3. Calculation of design effect (cluster sampling)

For cluster sampling, the resulting target sample size is adjusted for design effect, as outlined for example by Kish in 1965.³⁷ The adjustment is conducted by applying the following formula;

$$n_{eff} = n (1 + (M - 1) ICC)$$

Where:

n_{eff} = effective sample size

n = unadjusted sample size

M = average sample size per cluster or PSU

ICC = intra-cluster correlation

2. Interface

2.1 Sampling frame tab

➤ Mode of sampling

³⁶ Krejcie and Morgan (1970) "Determining Sample Size for Research Activities" (Educational and Psychological Measurement, 30, pp. 607-610)

³⁷ Kish, Leslie (1965). "Survey Sampling". New York: John Wiley & Sons, Inc

- Enter sample size: You can input the target sample size desired.
- Sample size based on population: The sample size will be based on confidence level, error margin and proportion.

➤ Type of sampling

- Random sampling from household list: Simple Random sampling, a random sample directly from the sampling frame, which consists of every unit in the population of interest, thus ensuring equal probability of each unit to be selected. See section 4 of the sampling guidance, under “simple random sampling”
- Random sampling from location list: Primary Sampling Units (PSUs) are randomly selected with probabilities based on population size to ensure equal probability of each sub unit (e.g. households). This option produces the target sample by PSU; the sampling of each sub unit (e.g. households) will need to be randomized in the field.
- Cluster sampling: Clusters, or Primary Sampling Units (PSUs), are randomly selected with probabilities based on population size to ensure equal probability of each sub unit (e.g. households). Each time a PSU is selected, a number of surveys defined by the “cluster size” set in the tool is attributed to the PSU. See section 4 of the sampling guidance, under “2 stage cluster sampling”
- Stratified: If ticked, a variable characterizing the stratification of the sampling will be asked to stratify the sample. It is possible to enter only one.

Illustration 1: screenshot of the sampling frame tab

The screenshot shows the 'Probability sampling tool' interface with the 'Sampling frame' tab selected. The 'Sampling type' section includes dropdowns for 'Mode of sampling' (set to 'Sample size based on population'), 'Type of sampling' (set to 'Simple random'), and 'Stratified?' (set to 'Not stratified'). Below these are input fields for 'Confidence level' (0.95), 'Error Margin' (0.05), 'Proportion' (0.5), and 'Buffer' (0.05). The 'Set up sampling frame' section features a 'Choose CSV File' button with a 'Browse...' option and 'No file selected' text, and an unchecked checkbox for 'Use test data'. An 'Apply' button is located at the bottom right.

➤ Parameters to be entered

- Confidence level: The desired confidence level.
- Error Margin: The desired error margin.
- Proportion: The expected proportion in the sample.
- Buffer: The desired percentage buffer.
- Cluster size: The minimum number of interviews to conduct by cluster (only for cluster sampling)
- ICC: Intra-cluster correlation: average value estimated in previous assessments = 0.06 (only for cluster sampling)

➤ Loading files

The app takes only files .csv; each row of the csv need to be the primary sampling unit. The headers of the dataset must NOT contain special characters. You can use some example data by ticking 'example data'. The .csv files needs to be separated with comma. Csv files saved with a French version of Excel will use semicolon for separation and will have to be modified prior to uploading.

When the files is loaded, depending of the type of sampling loaded before, some parameters will have to be specify:

- **Cluster:** The variable including the name of each cluster, there should be no duplicates in this column.
- **Stratification:** The variable in the data defining the stratification.
- **Population:** The variable in the data defining the population number by PSU. Must be > 0 and > to cluster size

➤ Sampling target

When all the above parameters are defined, under the target section will be computed the target based on the confidence level and margin error defined previously.

Click Apply to compute the target

2.2 Sampling Tab

When target appear in the tab, go to the 'Sampling tab' and click 'Sample!' to produce the sample. You can then download the list of units sampled and the sampling summary in Excel format (see example below)

Illustration 2: screenshot of the sampling tab

The screenshot displays the 'Sampling' tab of the 'Probability sampling tool'. It features a navigation bar with 'Introduction', 'Sampling frame', and 'Sampling' tabs. A 'Sample!' button is prominently displayed. Below this, the 'Sample summary' section contains a table with the following data:

Stratification	# surveys	# units to assess	Cluster size	Cluster size set	ICC	DESS	Effective sample	% buffer	Confidence level	Error margin	Population	Sampling type
Central	405	177						0.05	0.95	0.05	1142835	2 stages random -st1
Eastern	403	139						0.05	0.95	0.05	98871	2 stages random -st1
Western	404	142						0.05	0.95	0.05	186792	2 stages random -st1

Below the summary table, there is a 'Sample' section with a table listing individual units:

id_sampl	Survey	strata_id	Region	Zone_name	District_ID	District_name	VDC_name	VDC_code	P_CODE	HH_total	Ind_total
id_10	1	Central	Central	Narayani	31	Makwanpur	Bhaise	310006	C-NAR-31-006	1388	6717
id_100	3	Eastern	Eastern	Sagarmatha	12	Okhaldhunga	Shreechaub	120047	E-SAG-12-047	581	2856
id_101	1	Eastern	Eastern	Sagarmatha	12	Okhaldhunga	Singhadevi	120048	E-SAG-12-048	455	2099
id_102	5	Eastern	Eastern	Sagarmatha	12	Okhaldhunga	Balaku	120003	E-SAG-12-003	835	3987
id_103	2	Eastern	Eastern	Sagarmatha	12	Okhaldhunga	Taluwa	120050	E-SAG-12-050	447	1996

Annex 3: Example from REACH Jordan- A guide to GIS-based sampling for host community projects

Background

In the absence of a full, comprehensive GIS data set on household locations and population for Jordan, since 2016, REACH started utilizing a unique and detailed geographic method for developing randomized, weighted GPS sampling points for host community assessments in Jordan. The goal of this sampling approach is to ensure that *every household has an equal chance of being selected for a household interview*, while also recognizing that the baseline geographic data we have in-house is not comprehensive enough to allow for a simple selection method as we do in locations where we have household address, for example. As a result, we utilize a detailed methodology based on population density weighting and a randomized sampling loop to ensure that areas with higher population are weighted accurately against areas fewer people. This document outlines the process in detail, and can be replicated in other contexts with limited GIS data on population density and household locations.

Technical approach & summary of steps

The sampling approach requires various methods of geospatial analysis in GIS software, weighting and aggregation in MS Excel, and randomized sample generation using an R statistical package. Before beginning this process, you must very clearly understand the sampling requirements of the specific assessment, including, but not limited to:

- The sample locations and boundaries (e.g. municipalities, districts, village boundaries, etc.);
- Sample size per distinct sample site / location / cluster (i.e. geographic strata);
- Sampling constraints and / or additional stratifications around demographics or gender.

Once you have a clear definition of the above, you can begin preparing your framework and methodology around the available GIS data to conduct the population analysis and GPS point selection. You should proceed with the sampling process through the following steps, which are subsequently outlined in detail.

- i. Setup a workspace in ArcGIS to include baseline GIS data on the sample locations, villages, roads, and anything else you would need to make detailed maps of your areas of interest for the assessment;
- ii. Use the *HexagonalPolygon101* toolbox in ArcGIS (must install—can be found on server in Jordan) to create a grid of hexagons over your area of interest, at a size defined as appropriate for enumerators to use as a radius for identifying households in the field;
- iii. Remove hexagons from the base layer that are clearly outside of populated areas or areas where there are roads or access constraints due to security and or permission to assess in a specific area;
- iv. Using either a spatial join or an Inverse Distance Weighted (IDW) interpolation technique in GIS, you will need to give each hexagon / point a value based on the value of population within or around the hexagon;
 - This is the critical step and will vary depending on the available data to utilize as a proxy for population at a detailed enough level for differentiating within urban and peri-urban areas.
- v. Import the attribute table of the weighted hexagon shapefile into MS Excel and create a column for a probability index based off of the population “value” for each point;
- vi. Using R statistical and scripting software, use a pre-designed script to allow for selecting points based on a looping function and accounting for probability given in the MS Excel aggregation step above;
- vii. The output of the R script will provide a CSV of points selected by the ID from the hexagon attribute table. Create a summary pivot table with a column for ID and a column for number of times selected, and use this to join back to the shapefile in ArcGIS and create a new shapefile for only those hexagons selected in the sample;

- viii. Create maps for each sample location and a separate table showing point ID, number of interviews per point, geographic location, and GPS coordinates for usage on smartphones in the field;
- ix. Create a GPX file that can be loaded to mobile phones and used directly in OsmAnd maps and navigation application.

Step – by – step guide

Using the numbering system from above, the following is a detailed instructional guide to create a randomized GIS sample using population density and spatial analysis to allow for a representative sample.

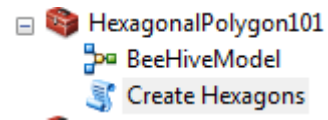
i. **Setup ArcGIS workspace**

Within ArcGIS, the workspace needs to be setup in a projected coordinate system to allow for distance measuring and calculations that will be used in a number of steps throughout this process. For Jordan, we use *WGS 84 UTM Zone 36N* or *Palestine 1923 Palestine Grid*, both of which are defined in meters.

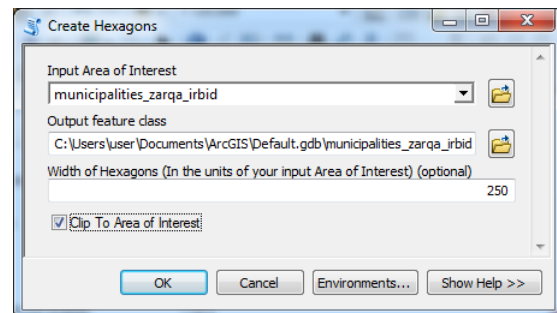
ii. **Create hexagonal grid for sampling**

To create a grid and set of points that enumerators can use in the field to identify households, we utilize the *HexagonalPolygon101* toolbox which is available to download online or on the Jordan server: *Z:\REACH\JOR\Resources\Software\ArcGIS Toolboxes\Hexagon101*. This tool creates a grid of hexagons in your area of interest and will be based upon pre-defined size.

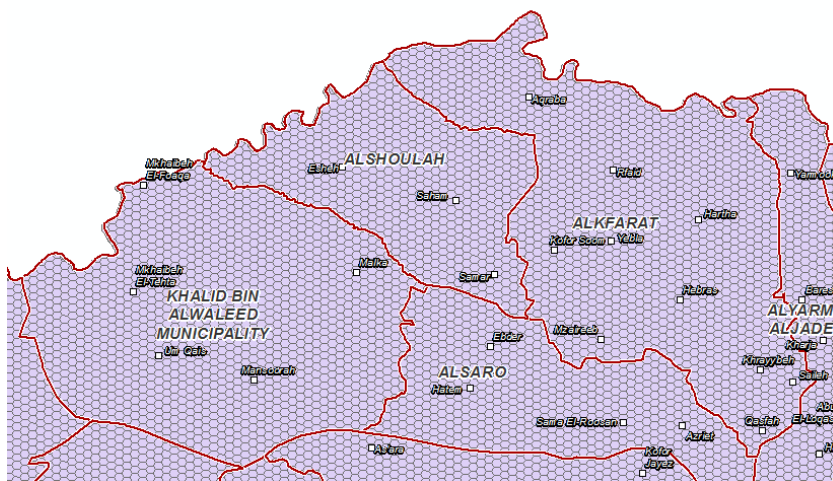
- a. Under the ArcToolbox, find the *HexagonalPolygon101* (you may need to add this to your workspace) and select “Create Hexagons”;



- b. In the dialog box, select the appropriate area of interest (i.e. your shapefile for all of the geographic areas where you will conduct the assessment), define an output feature class, and specify the appropriate distance for each hexagon diameter (e.g. 250), a check the box to clip the output feature class to the area of interest.



The output should look something like the below, and note that this process will take a very long time and might need to be left overnight;



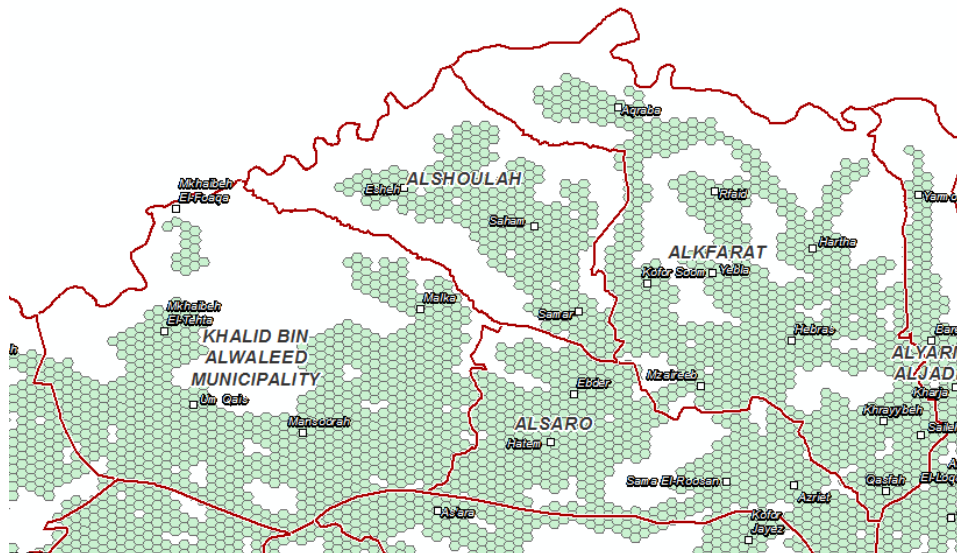
iii. **Filter points that are in uninhabited areas**

It is important to understand the full scope of this exercise and be aware of all tasks and final output throughout, in order to ensure that the process is as detailed as possible. This step is a good example of thinking towards the random sample step. Given that each hexagon in the shapefile will potentially be used in the sample base, it is useful to attempt to remove the points that are clearly in uninhabited areas and remove them from the master sample file. This can be achieved in a number of ways:

- a. Use Landscan gridded population data to remove hexagons that are close to uninhabited zones by conducting a spatial join of the hexagons to the Landscan population point shapefile ("Z:\REACH\JOR\Country_GIS\Common_Datasets\Landscan\jor_landscan_point.shp"). Each point represents a specific population within 1 sq/km—you can remove a significant number of hexagons that are closest to points that have a value of less than 10 people per sq/km;
- b. Remove hexagons that are further than 100m from roads. Be sure to use the latest export from OSM and ensure that you're not using tracks / trails / roads under construction (just actual roads);
- c. Remove hexagons that are within 1km from the border with Syria;
- d. Discuss specific locations with the field team that are known to be inaccessible such as universities, Palestinian refugee camps, military bases, factories, etc.

Note: it is important to ensure that all of your spatial data is in the same and correct projection for these steps and those following, as you will be measuring distance and creating files based on distance, you should be in either of the two listed above. Later you will switch to WGS84 to create the coordinates that will be used for the field maps and the smartphones for navigation in the field.

The outcome of these steps should give you something more filtered and naturally representative of populated areas, such as the below



iv. **Perform GIS analysis to assign hexagons with population values**

In this step, we will assign each hexagon a value based on population statistics. In Jordan, we have executed this on a number of times using two distinct methodologies based on data availability. Both

steps involve some form of interpolation or weighting based on GIS data on population; however the methods differ due to the wide difference between the data sets we have in-house. In one case, we have detailed data on the Jordan water network and the locations of water customers, which can be used as a proxy for population density at the household level. However, this is only for the northern governorates and limits this approach to only those areas. When we have assessments outside of those governorates we instead use Landscan gridded population, which is only estimated at the 1 sq/km scale and therefore much less detailed.

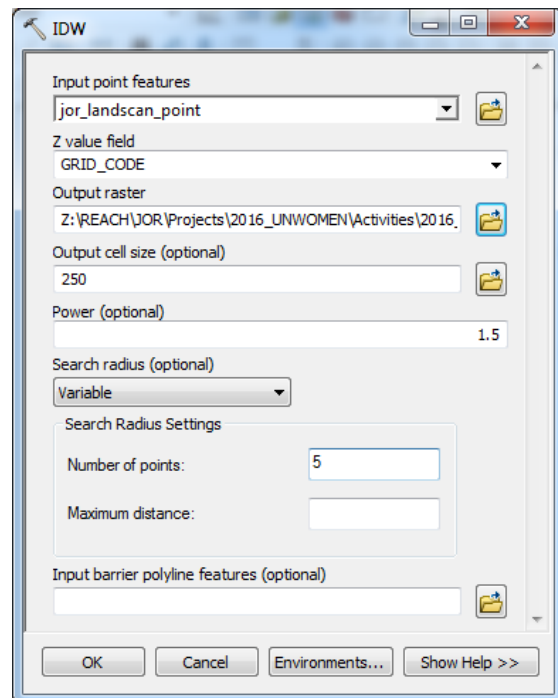
a. **Water customer GIS data analysis**

In the Country GIS folder on the Jordan server, there are a number of infrastructure related files that can be used in this analysis: Z:\REACH\JOR\Country_GIS\Common_Datasets\Infrastructure. The most important of these is the “customer.shp” data set which contains the point location of every customer of the water network, by number of people / households at that location. Using this data, you should conduct a spatial join calculating the total sum of customers that fall within each hexagon. This process will provide a population value column and be suitable for creating a weighting per sample site to use for drawing a randomized sample.

b. **Landscan population density interpolation**

Using the Landscan data is more nuanced and arbitrary than simply conducting a spatial join, but still achieves the goal of providing a realistic weight to each point on the grid to better represent human settlement across the country. The following steps are a general guide to this process

1. As shown in the previous step to reduce the number of hexagons in the base file, prepare the Landscan shapefile for analysis by creating a definition query that only keeps points that are more than 10 (# of people) in the GRID_CODE column, so it reduces the amount of data to be processed;
2. In ArcToolbox navigate to *Spatial analyst > Interpolation > IDW* where we will use an Inverse Distance Weighted interpolation method to create a surface raster of the same size to the hexagon grid, weighted using the Landscan points shapefile;
3. In the dialog box for the tool, put the Landscan points shapefile as the input points feature and specify the value field as “GRID_CODE”;
4. Specify that the output cell size is 250 (meaning 250m across, and again it’s required that the Landscan points shapefile be projected into one of the two projections listed above);
5. Put 1.5 as the value in the power field as this is the middle value on the scale that determines how much weight to assign to each point used in the interpolation;



6. Put 5 as the number of points to use in the analysis, although this can surely be changed for later sampling exercises depending on the need and the data source;
7. Convert the output raster to a point shapefile and conduct a spatial join to the hexagons layer to assign each hexagon the value of the point that falls within it, which as a result gives each hexagon the value of the IDW analysis. The surface raster created in the IDW and the resulting point shapefile should be the same scale as the hexagon grid, allowing for a seamless, one-to-one spatial join.

v. **Create a probability ranking for the grid points**

After you've created a finalized and cleaned shapefile of population-weighted hexagons, it's time to begin the analysis process and the final steps to prepare the data set for sample extraction. You will need to create a probability or weighting field, which is easiest using MS Excel. This will mean taking the attribute table from the hexagons shapefile and directly importing to Excel to perform the analysis. This table will then be used in the R script to draw the sample points using probability and the hexagon ID from the shapefile attributes (the later to ensure you can join back to the shapefile once the sample has been drawn).

- a. Before doing any of the extraction and analysis, you should take the existing hexagons shapefile (the one with the population analysis join) and re-project this to WGS84. Now that all of the distance calculations have been conducted, it's no longer necessary to maintain this projection and switching back to WGS84 will ensure that the coordinates align better with the smartphone navigation applications and that the field maps are in Mercator which aligns with the smartphones;
- b. Create LAT & LON columns in the shapefile and use the calculate geometry tool to generate coordinates for each hexagon in decimal degrees;
- c. Open the .dbf of the shapefile in MX Excel and create a new column called "*probability*" (see the screenshot below). In this example, the "*count_*" field is the population value given to that point during the spatial join to the water network GIS data file, and this field is used to generate a weight for the point against the total population within the strata of interest (i.e. municipality, district, community, etc);
- d. The "*probability*" column (column F) is based on a formula that divides the cell value in column E (e.g. E2) by the sum of column E for each unique location. This is done by using a SUMIF formula, demonstrated in cell F2 in the screenshot to the right.

	A	B	C	D	E	F	G
1	FID	hex_ID	OBJECTID	usaid_comm	count_	probability	
2	9325	9326	63	Al Salhiyyeh w Nayfeh	1	=E2/SUMIF(D:D,D2,E:E)	
3	9346	9347	148	Al Salhiyyeh w Nayfeh	1	0.000547345	
4	9347	9348	62	Al Salhiyyeh w Nayfeh	0.5	0.000273673	
5	9348	9349	61	Al Salhiyyeh w Nayfeh	0.5	0.000273673	
6	9349	9350	62	Al Salhiyyeh w Nayfeh	0.5	0.000273673	
7	9350	9351	61	Al Salhiyyeh w Nayfeh	6	0.003284072	
8	9351	9352	62	Al Salhiyyeh w Nayfeh	1	0.000547345	
9	9352	9353	63	Al Salhiyyeh w Nayfeh	1	0.000547345	
10	9353	9354	147	Al Salhiyyeh w Nayfeh	0.25	0.000136836	

vi. **Generate the random sample using an R script**

Now that the table has been created with a weight column, we can begin designing and customizing a script in R that will allow us to draw a random sample based on probability. This involves using a basic looping script that is based upon the *sample* function from the R statistical package. Please look carefully through this and the documentation available online to ensure that you can successfully tweak the script to fit your data and sampling needs.

	A	B
1	usaid_comm	sample_size
2	Al Salhiyyeh w Nayfeh	95
3	Al Taybah	96
4	Al Wasteyya	98
5	Hosha al Jadeeda	95
6	Khaled ibn Alwaleed	96
7	Mu'ath ibn Jabal	97
8	Noaimh	96
9	Sabha w al Dafyaneh	95
10	Um Ejmal	96

- a. Ensure that the sampling table has been saved as a CSV so it can be read into the R workspace as a text file and manipulated to perform the sampling;
- b. If you have different sample sizes for each site, you need to create a separate table that lists the sample sites and the number of samples per location (see screenshot to the right). You will see in the script outline that you can delete this step if the sample size is the same for each location;
- c. The script pasted below, and provided as an attachment to this document, outlines the process for drawing a random sample from the formatted CSVs.

```

1 setwd ("C:\\Users\\user\\Desktop\\usaid sampling temp\\script")
2 db <- read.csv("usaid_sampling_probability.csv",header=TRUE)
3 sam_tab <- read.csv("usaid_sample_size.csv",header=TRUE)
4 usaid_sample <- c("start_")
5 for (i in 1:10)
6 {
7   sam <- db[as.character(db[,4])==as.character(sam_tab[i,1]),]
8   for (j in 1:sam_tab[i,2])
9   {
10    samp_final <- sample(as.character(sam$hex_ID), 1, replace=FALSE, prob=sam$probability)
11
12    usaid_sample <- c(usaid_sample,samp_final)
13  }
14 }
15
16 write.csv(usaid_sample,"usaid_sample_final_.csv")

```

- d. To understand the above script, you must understand some basics of coding in R (or other coding languages). However, it's a very simple code about which you only need to grasp a few key concepts. Each section in the code is numbered above and explained below:
 1. Sets the working directory where you have your data saved and your output file will be saved. Follow this format when changing to be accurate to your workspace;
 2. Creates an object in the R workspace called "db" which reads in the CSV file titled "usaid_sampling_probability.csv" within the folder you designated as the workspace. header=TRUE is a component of the function read.csv where you indicate that your data set has a header (see online documentation for more detail);
 3. Creates an object in R called "sam_tab" that reads in the CSV for the sample size by location;
 4. Creates an object called "usaid_sample," a vector object that begins with "start_" as the first item. This serves to create an object that will act as an anchor and be used to join the selections from the sample loop, further down in the script;
 5. The beginning of a "for" loop in the R script, which is designed to repeat 10 times (10 being the number of sample locations in the assessment);
 6. Beginning at component 6 this will repeat itself 10 times per the loop design in the previous line of code. Each time this creates an object called "sam"

which takes the “db” object (remember, this is the object created above to host your raw data) and *subsets* (technical R procedure) the data when column 4 of “db” is equal to column 1 of “sam_tab.” Think about this like a filter process that, within the loop, extracts the data only when the value of “usaid_comm” is equal to the relevant value in the “sam_tab” data set;

7. This step creates a second loop within the larger loop, and is designed to repeat itself the number of times as specified in the second column of the “sam_tab” object (the table you created above for the number of samples per location). Simply, each time the larger loop runs (10 total) this step says repeat the number of times specific in column 2 of the “sam_tab” object (where the sample size values are hosted);

Note: what you now have is a master loop that repeats itself the number of times needed to draw a sample for each location. The master loop identifies the appropriate table rows for each location based on a join to an external sheet that has location key (it literally filters the master data one by one according to the key). The second component of the loop tells it to repeat the number of times specified in the sample size table (column 2) rather than the same number of times per location. If you have a sampling framework with the same number of locations, the second loop code can simply mimic the first (e.g. for (j in 1:96)). With this, instead of being unique for each location, it will repeat this second portion of the loop 96 times, and 10 times overall.

8. Within the second component of the loop, this portion creates an object called “samp_final” which utilizes the *sample* function within the R package. The sample function then reads in the data in the “sam” object from the first portion of the loop, draws from the column “hex_ID”, selects 1 record from the data, and accounts for the probability indicated in the column “probability” (you need to research the structure of this function to fully understand what it is doing, and the nature of subsetting or calling columns by using the \$ and the object / data name). All of the values in this formula need to be accurate to the objects and data columns they are referencing, otherwise it will break;
9. Within the same loop, this recreates an object called “usaid_sample” that is a join between the original “usaid_sample” object and the “samp_final” object generated in the above section of the loop. The result is that you have an object called “usaid_sample” that is continuously used by the loop to add a sampled record. In the R console you could test this by entering the “usaid_sample” object and see the resulting data set, which is the full sample;
10. Takes the final object of “usaid_sample” and writes an output CSV into the workspace.

	A	B
1	hex_id	sample_size
2	100	2
3	117	1
4	121	1
5	138	1
6	139	1
7	174	1
8	184	1
9	214	1

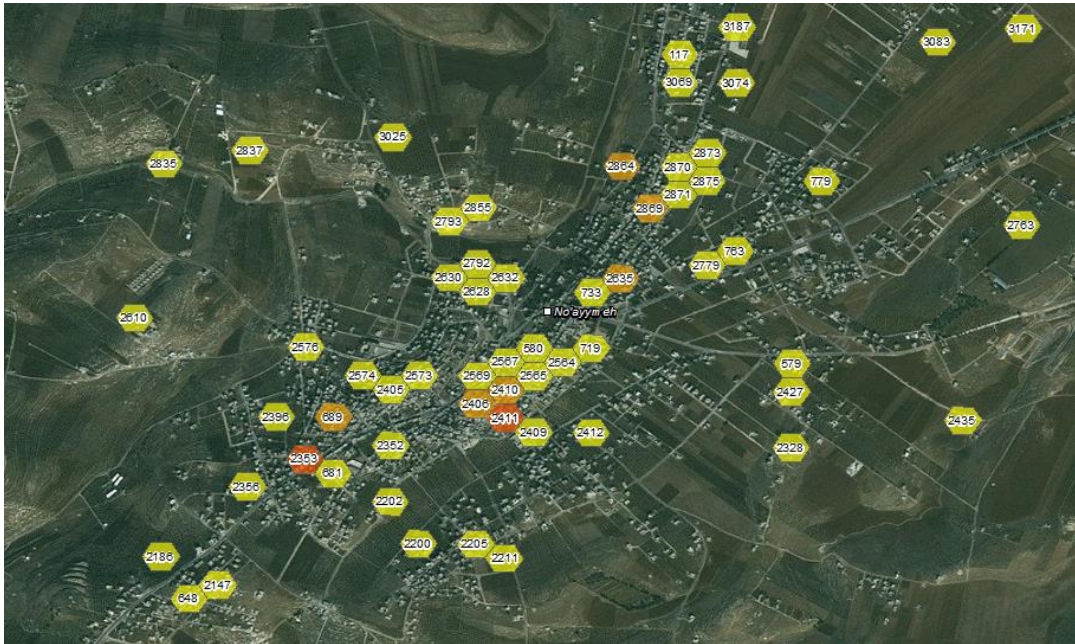
vii. **Create table of point IDs and sample size**

Using the output CSV sample from the R script, you will have a table that lists the hexagon IDs in the order that they were selected in the script. It’s very important to understand that through this script and sampling approach one ID can be selected multiple times, so before re-linking this to the shapefile we need to create a summary (pivot) table that indicates the number of times the ID was selected. The result should look like the example to the right, where the ID is listed as a row and the number of times it was selected in the next column, which you can then join back to the master shapefile in ArcGIS.

viii. **Produce maps and tables of sample points for field teams**

After performing an attribute join of the sample table to the master hexagon file (remember to select the option in the join window to ‘keep only matching records,’ which will ensure that only the hexagons in the sample appear on the map), you now have a layer showing only hexagons that have been selected in the sample, with a column in the table indicating the number of interviews that should be conducted within the area of the hexagon (*Reminder: at this point you should create a new MXD workspace in WGS 84 since you will now make maps for the field, and the master hexagon file at this point has now reverted back to WGS84 from being projected*).

- a) You can visualize the hexagons on a color scale indicating the number of assessments



per location, for ease of coordination in the field (see screenshot below);

- b) Each point should have an ID that can be clearly seen on the map, which will also be used in the field for team leaders to look at the map and reference a formatted table showing the location IDs, geographic information (sample location, district, governorate, etc), and GPS coordinates to use for navigation with smartphones;
- c) Put important base data on the maps like village names, admin boundaries, roads, and normally satellite imagery for ease of reference in the field.
- d) Create a separate table—using the attribute table of the final hexagon sample layer—which the teams will print and use in the field for reference during data collection. This table should include key information and cleanly organized so the teams can print the table and use in the field (see screenshot below).

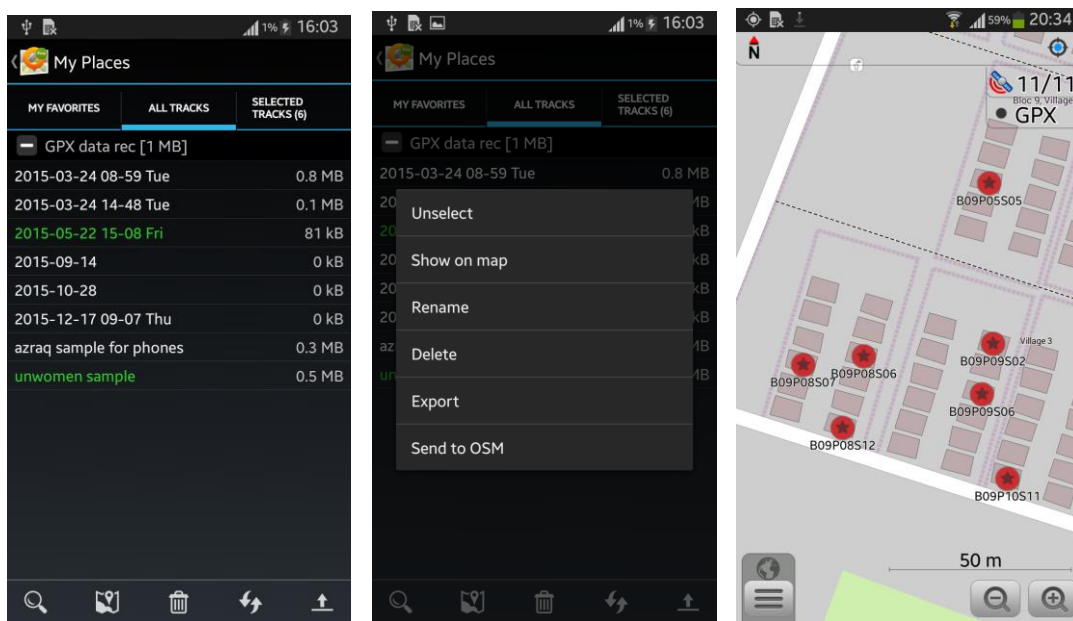
ID	Female	Male	Total Interviews	Name_AR	Name_EN	Gov_EN	Gov_AR	LON (E)	LAT (N)
2275	0	1	1	الزرقاء	AL ZARQA	AL ZARQA	الزرقاء	36.08525659	32.04474179
10162	0	1	1	الزرقاء	AL ZARQA	AL ZARQA	الزرقاء	36.08097051	32.07633426
10312	0	1	1	الزرقاء	AL ZARQA	AL ZARQA	الزرقاء	36.09698872	32.07284165
10314	2	0	2	الزرقاء	AL ZARQA	AL ZARQA	الزرقاء	36.08092787	32.07182556
10356	0	1	1	الزرقاء	AL ZARQA	AL ZARQA	الزرقاء	36.07401756	32.068491
10357	0	1	1	الزرقاء	AL ZARQA	AL ZARQA	الزرقاء	36.08088523	32.06731685
10359	1	0	1	الزرقاء	AL ZARQA	AL ZARQA	الزرقاء	36.09005622	32.06725364
10541	1	1	2	الزرقاء	AL ZARQA	AL ZARQA	الزرقاء	36.11793047	32.06758404

ix. **Create GPX file for smartphones and mapping applications**

Now that we have a fully generated sample with accompanying maps and tables for the field teams, we can go a further step and create a GPX file that can be loaded to the phones and visualized in any offline mapping application (allowing for easier navigation and identification of the point in the field)—in Jordan we use OsmAnd which relies upon OpenStreetMap data and is consistent with the

field maps we produce. For this step, you can use the sample table you have already produced; however, to ensure that the points will display a label on the phone, you need to rename the ID field to “Name” so the mapping application can recognize this field and display the appropriate point ID.

- One you have a formatted table with a “Name” column, you can use QGIS to read-in the CSV and save the file as a GPX file format (<http://gis.stackexchange.com/questions/157113/how-to-prepare-and-upload-points-lines-to-gps-using-qgis-2-8>);
- Next, plug-in the smartphone to your computer (make sure that the phone has OsmAnd installed) and find the osmand folder. You can simply copy/paste the GPX file into the folder named ‘tracks’, then you will be able to display the file by finding the file under “My Tracks” in the settings / preferences menu (http://osmand.net/help/faq.html#importing_gpx) – see screenshots below for this step and the final visualization in the OsmAnd application.



Additional notes

- It is highly recommended that, in addition to the process outlined in this document, you create a series of buffer, or top-up, sample points that can be used in the field in the event that the original locations are infeasible for household interviews. This is based on previous assessments and lessons-learned, due to the fact that this sampling method may, at times, send teams to locations where there are no households or locations that are not permissible for assessments;
 - The most efficient way to create buffer points is to adjust the R script and create an appropriate amount of top-up points per sample location (e.g. change the loop repeat amount to something like 20 and create a separate output CSV with “_buffer” at the end of the name);
 - Once the buffer is generated, create a similar table as with the original sample, however with this you should clearly include the *number showing the order that the point was selected in the sampling loop*. The output from the R script will list the order in which each point was selected (and it will automatically be ordered by sample location) and this should be made clear in the buffer table since the teams need to select the buffer point in this order to maintain the randomness of the sample (otherwise it’s likely that they would just choose the point that is closest to where there are in the field at that point in time).

Annex 4: Sampling considerations for remote, phone-based data collection

a. Sampling prerequisites for phone-based data collection: Overall

- ✓ **Availability of phones** (and specifically smartphones if using chatbot methods) among the population of interest
- ✓ **Decent phone network coverage for intended respondents** in the areas of interest
 - Map out phone network availability as early as possible, with such mapping done at the unit that you will be considering “strata”
 - The GSMA publishes network coverage maps globally (available [here](#)) based on their members’ latest network coverage data.
- ✓ **Decent phone network coverage for enumerators** (either at home or place of data collection)
- ✓ **Sufficient airtime on different networks** to ensure you can reach respondents that use different mobile networks with wide coverage
- ✓ **Ability to set up the required data protection measures** and take all the steps to ensure personal and sensitive data (e.g. names and contact details) are securely managed within the team (see [IMPACT SOPs for Management of Personally Identifiable Information](#))

b. Sampling prerequisites for phone-based data collection: Probability (random) sampling

- ✓ **Availability of reliable, anonymised and comprehensive lists reflecting the sampling frame** to ensure equal probability of selection for all units within population of interest
 - What is it: Using existing lists (e.g. consolidated, anonymised beneficiary lists from humanitarian partners with phone numbers) to enable randomised list-based selection
 - Pros: If available and contains the required information, the most straightforward way to randomise respondent selection
 - Cons: Unavoidable sampling bias towards those with a functioning phone and in a phone network area; difficult to apply stratification if additional metadata (e.g. location, displacement status) unavailable
- ✓ **[If list is not available] Ability to apply random digit dialing (RDD) techniques or other approaches to construct the list and the required sampling frame**
 - What is RDD: RDD is a technique used to draw a sample of households by using a randomly generated telephone number as a link to the respondent. In other words: the population of interest consists of all possible phone numbers in the area and all of these numbers have an equal probability of selection. For an example of how we have used a sampling approach like this in the past, please see pages 3-4 of [this report](#).
 - Pros: Enables random sampling even if Option 1 above is not feasible
 - Cons: Non-existent/ unassigned numbers sampled; higher chance of non-response; for household surveys, potential duplication of households (although this should be possible to mitigate by checking at the start if the household has already been interviewed either by simply asking or verifying against a household unique identifier such as refugee registration ID); difficult to implement any type of stratification (more than ten calls could be made till you find the respondent profile you are looking for)

c. Prerequisites for phone-based data collection: Non-probability (purposive or snowball) sampling³⁸

- ✓ **Availability of reliable key informant (KI) networks and contact details** to gather the required type of information
- ✓ **Ability to diversify KI profiles** as needed/ based on research objectives e.g. both males and females; different age groups; minorities or vulnerable demographic groups; etc.
- ✓ **Ability to ensure as wide a coverage of the population of interest as possible** using existing KI networks (for e.g. snowballing to ensure most settlements within a district are covered)

³⁸ In some cases, it might also be worth considering the feasibility of respondent driven sampling (RDS) which is essentially a variation of snowball sampling which uses social network theory to overcome the respondent bias limitations associated with snowball sampling. Specifically, RDS uses information about the social networks of participants recruited to determine the probability of each participant’s selection and mitigate the biases associated with under sampling or over sampling specific groups. For more on RDS, see also: WHO & UNAIDS; [‘Introduction to HIV/ AIDS and sexually transmitted infection surveillance Module 4](#) Unit 1: Introduction to respondent-driven sampling’ (2013), p.17-25

Annex 5: Troubleshooting issues encountered due to inaccurate information in sampling frame

1. Problem statement

MSNAs are usually based on household level interviews selected with probability sampling strategies.

Given the lack of available household lists in the countries of intervention, the sampling strategies usually rely on a list of locations defined as polygon or a point with an estimated number of households sampled with two stage random sampling or cluster sampling strategies. The sampling is usually done with replacement in order to have a self-weighted survey at the stratification level. The use of these methods tend to create some difficulties, when the sampling frame is not accurate especially in the following situations:

1. New location or sampling unit to be added in the sampling frame after start of data collection.
2. Population targeted not present in the location assessed, or change in access during data collection.

2. Adding unit(s) to the sampling frame

The need to add sampling frame units arises when a location that was not accessible during the research design becomes accessible in the course of data collection, or when a population group appears to be present in a location but it is not in the sampling frame.

Examples: Some restrictions are lifted regarding movement of humanitarians in some the assessment areas, and the team wants to include the location in the sampling; A team is going to a location to interview host population – there should be only host according to the sampling frame – but the team finds some refugee population as well (also part of the population of interest);

In that situation, the following questioning will help to take the right corrective action:

- Could I get the refugee population size in the location from a reliable source?
- Can I handle PSU level weighting during analysis?

If **NO** to any of the previous questions: then we do not attempt to interview population that were not part of the initial sampling frame. The presence of refugees in this particular location is reported in the limitations of the methodology section.

If **YES**, to all the questions follow the next steps; some adjustments are possible to include to the population in the sampling frame.

2.1 Getting missing population number(s) from a trusted source

Adjusting the sampling frame requires to have accurate (enough) numbers for the population. This is crucial to calculate the probability of inclusion of a household and to calculate the PSU weights. It is also needed to correct the change in the probability of inclusion initially calculated prior to the addition of this new unit.

2.2 Calculate the number of surveys needed in this location

A simple way to determine the number of surveys needed for the additional location, is to calculate the relative size of the unit, compare to the total population inside the stratum, and based on that to proportionally assign the number of surveys (see example). Although this method corrects partially the difference inclusion, there is still a need to apply PSUs weight at the analysis stage (See next section)

Example: In table 1 below, Gundu VDC, became accessible. In order to include it into the sampling, we calculate its relative size to the district population, and multiply this by the number of survey in the district:

(Population of Gundu / Total size in district) * Sample size = $1257 / 68557 * 451 = 8.3 \rightarrow 9$ surveys needed

Table1: addition of PSU in the sample

VDC name	P Code	Total number of HH	Probability of inclusion	Survey buffer	Access
Sirutar	C-BAG-26-016	1033	0.015	12	OK
Duwakot	C-BAG-26-008	2412	0.035	24	OK
Gundu	C-BAG-26-009	1257	0.018	?	OK
Kautunje	C-BAG-26-011	4692	0.068	33	OK
MadhyapurThimiN.P	C-BAG-26-012	20302	0.296	136	OK
Nankhel	C-BAG-26-014	1225	0.018	10	OK
Sipadol	C-BAG-26-015	2278	0.033	16	Unsure
Sudal	C-BAG-26-017	1562	0.023	9	Unsure
Tathali	C-BAG-26-018	1264	0.018	7	OK
Bageswari	C-BAG-26-001	1137	0.017	5	OK
Jhaukhel	C-BAG-26-010	1631	0.024	9	Unsure
Nagarkot	C-BAG-26-013	973	0.014	11	Unsure
Balkot	C-BAG-26-002	3999	0.058	26	OK
BhaktapurN.P.	C-BAG-26-003	17639	0.257	108	OK
Changunarayan	C-BAG-26-004	1374	0.020	7	OK
Chhaling	C-BAG-26-005	1817	0.027	10	OK
Chitapol	C-BAG-26-006	1274	0.019	5	OK
Dadhikot	C-BAG-26-007	2688	0.039	23	Unsure
TOTAL		68 557		451	

3. Population targeted not present in the location assessed, or change in access during data collection

Inaccuracy in sampling frame, especially location that were planned cannot be accessed, can lead to the need of adding some surveys to the sample to reach the desired precision on the findings. In order to get a self-weighted survey, IMPACT usually uses sampling with replacement of the PSU. Adding survey after data collection, apply either to go back to the locations already assessed or to use PSU level weighting to account for the change in probabilities of inclusion. The next section describes some simple strategies to prevent the need for resampling on the go, and when these fail what to do to ensure that the bias created is corrected or properly accounted for.

3.1 Sampling design- Buffer

Plan a buffer that is big enough to ensure that the change in access to some location can be compensated by the additional surveys that are done in another location.

Dimensioning the buffer: a way to think about it is the number of surveys and / or locations from you sampling frame than you can drop without having to resample. In case where access in some areas remains uncertain - the buffer should be bigger, calculated based on the size of area likely to lose access during data collection. The buffer size can be different for each stratum.

Table2: example of sample with access assessment

VDC name	P_CODE	Total number of HH	Probability of inclusion	Survey buffer	Access
Sirutar	C-BAG-26-016	1033	0.015	12	OK

Duwakot	C-BAG-26-008	2412	0.035	24	OK
Gundu	C-BAG-26-009	1257	0.018	9	OK
Kautunje	C-BAG-26-011	4692	0.068	33	OK
MadhyapurThimiN.P	C-BAG-26-012	20302	0.296	136	OK
Nankhel	C-BAG-26-014	1225	0.018	10	OK
Sipadol	C-BAG-26-015	2278	0.033	16	Unsure
Sudal	C-BAG-26-017	1562	0.023	9	Unsure
Tathali	C-BAG-26-018	1264	0.018	7	OK
Bageswari	C-BAG-26-001	1137	0.017	5	OK
Jhaukhel	C-BAG-26-010	1631	0.024	9	Unsure
Nagarkot	C-BAG-26-013	973	0.014	11	Unsure
Balkot	C-BAG-26-002	3999	0.058	26	OK
BhaktapurN.P.	C-BAG-26-003	17639	0.257	108	OK
Changunarayan	C-BAG-26-004	1374	0.020	7	OK
Chhaling	C-BAG-26-005	1817	0.027	10	OK
Chitapol	C-BAG-26-006	1274	0.019	5	OK
Dadhikot	C-BAG-26-007	2688	0.039	23	Unsure
TOTAL		68 557		460	

Example: the teams takes a 20% buffer. Target sample size is 383 (95% confidence and 5% error margin) and we have a sample size of 385 and 20% buffer, 460 surveys.

In the table above the access has be characterized based on the likelihood of access. % locations were characterized with an unsure access – which is 16 + 9 + 9 + 11 + 23 = 68 surveys. With the buffer planned we can, if the access changes, drop the totality of the 5 locations: 460 – 68 = 392 and still have a representative sample with 95/5 precision level.

3.2 During data collection planning

If access due to security situation is likely to change in some areas, if the security permits, consider starting with the area where the access might change during data collection time.

Example: If the security is likely to change in Dadhikot and Nagarkot district (see table 1), preventing the team to physically access these areas for data collection, assuming it is currently safe for them, the field team could start by this area in order to minimize the risk of having to drop this location entirely due to access restrictions.

3.3 During data collection

Although the previous precautions were taken to ensure that the sample size target is reached, it is possible that an unforeseen change of access led to underachieving against the planned sample.

➤ Top-up with replacement

The country team has the capacity to go back to the site already assessed: In this case, additional PSUs will be redrawn from the original sampling frame (without the inaccessible areas). Because the sampling is done on all the units of the stratum, it is very likely that location already selected in the first sample will be redrawn again. In that situation the sampling strategy is kept (sampling with replacement) and there are no impact on the analysis.

Example: A team wants to compensate the deficit of 25 surveys due the inaccessibility of a location. It is possible to resample using IMPACT sampling tool³⁹:

- Load the sampling frame with inaccessible area removed, for a single stratum.
- Mode of sampling = “enter sample size”
- enter the number of survey that need to be added, 25 in this example.
- If cluster sampling – the cluster size needs to be similar than the initial sampling.

Illustration 1: screenshot of Impact Initiatives sampling tool.

The screenshot shows the 'Probability sampling tool' interface with the following configuration:

- Mode of sampling:** Enter sample size (dropdown menu is open showing options: Enter sample size, Sample size based on population)
- Type of sampling:** Cluster sampling (dropdown menu)
- Stratified?:** Not stratified (dropdown menu)
- Enter target sample size:** 0 (input field)
- Buffer:** 0.05 (input field)
- Cluster Sampling:**
 - Cluster size:** 5 (input field)
 - ICC:** 0.06 (input field)

➤ Top-up without replacement

The country team does NOT have the capacity to go back to the site already assessed. In that case, it is possible to draw additional locations from the sampling frame. However, the top-up sample will have an impact on the PSUs' probability of selection and by consequence will impede the self-weighting characteristic of the survey.

➤ Add location into the sample

To identify a new location to replace the one that was not accessible, you can use the sampling tool to recreate a sample with the same characteristics. Compare the new sample produced with the original one, and randomly select locations that were not accessed until you select enough survey to replace the location that was dropped.

³⁹ Available at https://impact-initiatives.shinyapps.io/r_sampling_tool_v2/

3.4 During analysis

➤ Applying PSU level weighting

When the survey is self-weighted for a given stratum, all units (e.g. households) in the stratum have the same probability of inclusion; there is no need to include weights to correct inclusion. In this situation, the use of weight is only needed to aggregate the data at a higher level than the stratum, for example at crisis level for a population group.

In the situations described in the previous section, the inclusion in the sample, after the sampling, of new units has compromised the self-weighted characteristics of the survey design. There is thus a need to adjust the probability of inclusions between the household in a same stratum based on the size of the PSU they are from. During the analysis, the survey will carry a PSU weight that will allow aggregation at stratum level based on their relative size.

Example: In the table below, the inclusion of a new VDC, Gundu.

Table2: PSU level weights

VDC name	P_CODE	Total number of HH	Probability of inclusion	Survey buffer	Access	PSU weights
Sirutar	C-BAG-26-016	1033	0.015	12	OK	0.5776
Duwakot	C-BAG-26-008	2412	0.035	24	OK	0.6743
Gundu	C-BAG-26-009	1257	0.018	9	OK	0.9371
Kautunje	C-BAG-26-011	4692	0.068	33	OK	0.9540
MadhyapurThimiN.P.	C-BAG-26-012	20302	0.296	136	OK	1.0016
Nankhel	C-BAG-26-014	1225	0.018	10	OK	0.8219
Sipadol	C-BAG-26-015	2278	0.033	16	Unsure	0.9553
Sudal	C-BAG-26-017	1562	0.023	9	Unsure	1.1645
Tathali	C-BAG-26-018	1264	0.018	7	OK	1.2116
Bageswari	C-BAG-26-001	1137	0.017	5	OK	1.5258
Jhaukhel	C-BAG-26-010	1631	0.024	9	Unsure	1.2160
Nagarkot	C-BAG-26-013	973	0.014	11	Unsure	0.5935
Balkot	C-BAG-26-002	3999	0.058	26	OK	1.0320
BhaktapurN.P.	C-BAG-26-003	17639	0.257	108	OK	1.0959
Changunarayan	C-BAG-26-004	1374	0.020	7	OK	1.3170
Chhaling	C-BAG-26-005	1817	0.027	10	OK	1.2192
Chitapol	C-BAG-26-006	1274	0.019	5	OK	1.7096
Dadhikot	C-BAG-26-007	2688	0.039	23	Unsure	0.7842
TOTAL		68 557		460		

Figure 6: Decision tree in case of uncertain access



Annex 6: Additional reading materials

1. Abdul Latif Jameel Poverty Action Lab (J-PAL); Online resources on Research Design (available [here](#))
2. ACAPS; [Humanitarian Needs Assessment: The Good Enough Guide](#) (2014)
3. Alexander, Jessica and John Cosgrave(ALNAP); [Representative sampling in humanitarian evaluation](#) (February 2014)
4. Babbie, E.; 'Survey Research Methods' (Second Edition, 1990)
5. Brown et al; '[GSR Quota Sampling Guidance](#): What to consider when choosing between quota samples and probability-based designs' (UK Statistics Authority, 2017)
6. Creswell, John W.; 'Research Design: Qualitative, Quantitative and Mixed Methods Approaches' (Third Edition, 2009).
7. Duflo, Esther; Rachel Glennerster and Michael Kremer; [Using Randomisation in Development Economics Research: A Toolkit](#) (2007)
8. ICRC; [Acquiring and Analysing Data in Support of Evidence-based Decisions](#) (2017)
9. ICRC & IFRC; [Guidelines for assessments in emergencies](#) (2008)
10. Imbens, Guido; [Experimental Design for Unit and Cluster Randomised Trials](#) (2011)
11. IMPACT Initiatives; '[Area-based Assessment with Key Informants: A Practical Guide](#)' (2018)
12. IMPACT Initiatives; [Memo on Cluster Sampling](#) (2020)
13. IMPACT Initiatives; [Standard Operational Procedure for Management of Personally Identifiable Information](#) (2019)
14. International Labour Organisation (ILO); [School-to-work Transition Survey: A methodological guide; Module 3: Sampling methodology](#) (2009)
15. Inter-agency Standing Committee (IASC); [Coordinated Assessments in Humanitarian Crises](#) (2012)
16. Skinner, Chris J.; "Probability Proportion to Size (PPS) Sampling"; [Wiley Stats Ref: Statistics Reference Online](#) (August 2016).
17. Stukel, Diana and Gregg Friedman; [Sampling Guide for Beneficiary-based Surveys](#) (2016)
18. UNHCR; [Needs Assessment Handbook](#) (2017)
19. UNHCR; [Sampling Decision Assistant](#) (online tool)
20. UNICEF; [Sampling Methods and Sample Size Calculation for the SMART Methodology](#) (2012)
21. U.S. Centre for Disease Control; Reproductive Health Assessment Toolkit for Conflict-affected Women (2007); [Chapter 3- Sampling Instructions](#), p.12-18
22. World Food Programme; [Sampling Guidelines for Vulnerability Analysis](#) (2004)
23. World Health Organisation (WHO), '[Steps in applying Probability Proportional to Size \(PPS\) and calculating Basic Probability Weights](#)'
24. WHO & UNAIDS; '[Introduction to HIV/ AIDS and sexually transmitted infection surveillance Module 4](#) Unit 1: Introduction to respondent-driven sampling' (2013), p.17-25