ALNAP Discussion Series
Improving the quality of EHA evidence

Method Note 1
February 2014

Representative sampling in humanitarian evaluation

by Jessica Alexander and John Cosgrave

Evaluations are meant to provide trustworthy feedback on a program - for real-time evaluations this feedback is crucial to being able to make course corrections, for ex-post evaluations this feedback can demonstrate effectiveness and accountability to both donors and affected people, and inform future programs. As evaluation professionals, we aim to collect data that is representative of a population at a certain time, and which can be used to compare with other snapshots in time (for impact assessment or trend analysis). If the information is to be used for advocacy purposes, this is especially important as entire studies have been discounted because the sampling framework was not deemed to be representative.

The **Evaluating Humanitarian Action Guide** (see Section 5.2.2, 120 - 124, for discussion on sampling) states that **random sampling** is critical if you want to generalise the results from the sample to the whole population. This kind of sampling means using random or quasi random options to select the sample and then employing a statistical generalisation process to draw inferences about that population. Representative sampling depends on having a sampling frame, a list of the whole population that the random sample can be drawn from. Generally we don't have a full list, but have some approximation of the full list, such as beneficiary lists, that we can draw the sample from.

The literature discusses other sampling options as well: **purposive sampling** refers to instances where researchers select cases with a particular purpose or goal in mind. These cases are typically rich in information to make analytical inferences about the population – such as community leaders, or people who have extensive experience with the population you're examining such as mothers and teachers if you're studying behaviour of children. But this leads to inevitable biases: your results may be skewed by people who may have strong feelings towards or against a project or those who have the time and resources to participate in an interview. Researchers such as Daniel (See: Guide on Sampling, 2013, Chapter 3) suggest that purposive sampling may be particularly appropriate where:

- the research is exploratory;
- there is a need for a quick decision or to target specific individuals;

- the desire is to provide illustrative examples;
- access is difficult or the population is very disperse;
- time and money are limited but you have access to skilled and highly trained personnel;
- a sampling frame is not available;
- you are using qualitative methods;
- you need to use easy operational procedures;
- you have a small target size.

Convenience sampling is the method with the most bias and should be avoided if possible. This approach uses samples which are readily available – such as a community closest to the side of the road, or families who are available to speak during the window of time that you have - and which may not allow credible inference about the population.

Probability/random sampling is the best method for generalizing findings to the rest of the population, but given the many constraints faced in humanitarian settings, this may not be possible and thus bias may occur. The ACAPS Technical Brief explains a biased sample as one in which all members of the population are not equally likely to be represented. Bias may occur because of under-coverage of some groups, due to large non- response rates among particular groups or because of lack of access. An example of bias would be an underestimation of income levels because those working longer hours in the sampled population have a higher non-response rate.[1] Certainly challenges in obtaining representativeness may be overcome: enumerators return to homes where someone is not there instead of moving to the next house where circumstances may be different; they work at different times of the day and during weekends to maximize reach to a representative group; they may utilise multistage sampling or source triangulation (triangulating data from several different perspectives, in order to check the validity of information and improve accuracy as a counter to the lack of extensive sampling). However, are there techniques that allow for more precise representativeness and rigor?

"Evaluators must decide whether it is possible to conduct a quality evaluation under the real-world constraints, and to select the strongest possible design and explain why certain choices were made."

As the demands for more rigorous and evidence informed evaluation (be it process, outcome or impact) increase, and donors are more eager to understand the results of their programmes, evaluators need to ensure more robust methodologies and precise sampling methods so that generalisations can be made. Evaluators must decide whether it is possible to conduct a quality evaluation under the real-world constraints, and to select the strongest possible design and explain why certain choices were made or not with regard to not only the sampling but also the evaluation method, approach and implementation given the budget, time and logistical limitations.

Cluster sampling is a common technique to obtain a random sample. A cluster sample is one where the total population is divided into groups or clusters and a simple random sample of the groups is selected. The benefit to cluster sampling is that it reduces the total number of interviews needed and provides more accurate results when most of the variation in the population is within the groups, not

1 ACAPS Technical Brief: How sure are you? Judging quality and usability of data collected during rapid needs assessments. August, 2013.

between them. But cluster sampling can be technically challenging and some evaluations have been criticised for attempting to conduct the technique without adequately describing how it was done, what it meant in that context and what implications this had for findings. Some agencies, such as ACF, have looked critically at their own evaluations and highlighted those where it was clear that the terminology used did not match the methodology applied.

This difficulty is not unique to evaluations. Nutritional surveys which have tried to use the same techniques ran into similar shortcomings. The CDC evaluated the monitoring of projects and the measurement of nutritional status and mortality in Somalia from the period 1991–93. The researchers found that the range of methodologies employed and outcomes measured were so variable and of such poor quality that they prevented widespread comparisons, and that, regardless of consistency, much of the data were simply not credible due to poor collection methods. In another review by the CDC in Ethiopia between 1999 and 2000, only 67 of the 125 surveys attempted to conduct a sample that represented the population served. Only 9 of those 67 surveys assigned clusters to the population in a manner that was proportional to the sub-units of the population and only 6 of those possessed the minimum number of clusters (30) and children (900) suggested by most nutritional manuals.

The 'spin the pen' approach is the most common technique for cluster samples: all dwellings from the centre to the edge of the cluster in the chosen direction are counted, one is chosen at random and interviews are conducted. Additional houses are selected along the line away from the centre. If the cluster edge is reached before the sample size is achieved, the interviewers move clockwise to the next house and back towards the centre conducting interviews along the way. Two biases have been noted with this approach – one is over-sampling houses close to the centre, the other is 'pocketing' - uneven spatial distribution of the variables of interest.[2]

"Survey teams found the GPS method and the grid approach easier to implement than 'spin-the-pen'."

Shanon et al have compared the 'spin-the-pen' approach with two other methods: one which superimposes a grid on a map of the cluster, randomly chose coordinates on the grid, and identifies the closest compound (houses in the setting tended to be in walled compounds); and the second which uses Global Positioning Systems (GPS) and satellite and aerial photographs to identify a randomly chosen point and the nearest compound to the right when facing north at the point. Survey teams found these new methods easier to implement than 'spin-the-pen'. They were most enthusiastic about the GPS method, although the grid approach was fastest. However, both alternative methods led to higher probabilities of choosing households in low density areas of the clusters.[3]

This approach overcomes some other problems – first, it reduces the work for interviewers, minimizes their discretion in choosing buildings, allows random selection with known probabilities, and minimizes 'pocketing' within clusters by spreading out the sample within the cluster. Unlike many previous techniques, it incorporates population (household) density, which permits calculation of

2 http://www.ete-online.com/content/9/1/5
3 Harry S Shannon, Royce Hutson, Athena Kolbe, Bernadette Stringer, Ted Haines. "Choosing a survey sample when data on the population are limited: a method using Global Positioning Systems and aerial and satellite photographs." *Emerging Themes in Epidemiology* 2012.

correct sampling probabilities. Enumeration of buildings is needed for only very small areas, a task that can be done before going into the field; and interviewers only need to enumerate households for multi-residential buildings.

Although this is a promising approach, there are considerations for evaluators. Satellite or aerial photographs for the GPS locations may be out of date and so confirmation of buildings should be done before visiting. Recent photos of sufficient resolution will be necessary to discern between buildings in dense areas. When resolution is poor, maps should be cross-referenced for accuracy in detail. To allow for the possibility that buildings chosen were not residential, 2nd or 3rd choice buildings within the cluster circle could be chosen to limit discretionary decisions.

The advent of mobile technology has led some agencies such as the Red Cross to use crowd sourcing as a way to obtain a representative sample. However, it isn't always certain whether people contributing data with their cell phone really represent those who are most in need and not just those who can access technology, making this more like a convenience sample.[4] More recently Ground Truth has used mobile phones to reach people in three Haitian camps to conduct a survey. Agencies collected phone numbers for all camp residents when setting up the program and numbers were then randomly chosen to call for an interview.

"Many evaluations are still using old approaches."

While these are promising trends, many evaluations are still using old approaches. Although they acknowledge the sampling shortcomings in the limitation section, these studies typically go no further in addressing what this means for the findings. It should be noted in these reports that a handful of cases can be valuable for illustrating processes or behaviours, it is rarely appropriate to make statements such as "Most children felt that …" or "Most vendors were …". Yet, many evaluations still do that.

For example, in Save the Children's Evaluation of their Somalia Crisis Response (April 2011 – 2012) the authors acknowledge this by stating "It is important to note that compared to the size of programme, the community sample interviewed is small and this therefore gives only an indicative steer," but findings are still presented as though they count for the programme population.

The above has detailed quantitative sampling methods, but representative sampling is often not appropriate for the qualitative approaches that predominate in much of humanitarian evaluation. A requirement for qualitative methods is "theoretical saturation" which is when the addition of further data yields no extra information to the properties of the categories already developed. Although this concept was first developed in Ground Theory approaches, it is valid across a whole range of qualitative approaches. Thus with Key Informant interviews or Focus Group Discussions you know that you have enough data when fresh interviews and groups are only reinforcing what you know already without adding anything new (Biernacki and Waldorf, 1981).

Qualitative approaches are typically characterised by small-N. For these cases, snowball sampling is a good method, particularly to find key informant interviews. This method uses chain referral sampling, where each interviewee is asked to suggest the names of other interviewees who can

4 http://www.redcross.int/EN/mag/magazine2012_1/18-19.html

speak authoritatively on the topic. The chain continues until no new names emerge and can lead to dozens of interviews. In these cases we almost shy away from representative samples, as we purposefully seek out people who can provide the most amount of information. If we want to know what factors contributed to success in a microcredit project, for example, using a qualitative approach will bias our sample by interviewing the most successful users of credit and asking them what led to their success. This means that we cannot generalise to the whole population, but we can say something about success factors.

However, this still leaves the small-N researcher with many sampling decisions, including the issue of sample size. Normally the indication of adequacy for small-N samples is that of "theoretical saturation" or "data saturation" or "information redundancy". This is the point at which no new data or themes emerge from the data. However, there is very little operation guidance on how many interviews are needed to reach saturation. In 2011, Creswell and Clark (p. 174) suggested four to six cases for a case study and 20-30 interviews for a grounded theory study. However, these seem to be rules of thumb based on practice.

For focus groups, Morgan (1997) advises from experience that few new themes emerge after the third focus groups. Unlike these number that seem to be based on experience, a study of data saturation in key informant interviews (Guest et al., 2006) concluded that data saturation had been reached by six to twelve interviews.

However, it should be clear that in both cases Morgan and Guest are talking about homogeneous groups. Most humanitarian action has a variety of stakeholders each of which should have their voice heard. Thus if you were looking at education and youth, you might have three focus groups each with male youth, female youth, teachers, and parents. Similarly you might need a series of interviews with local community leaders, women's group representatives, NGO field staff, NGO managers and so on.

"Whether using large-N or small-N methods, humanitarian evaluators should explain their sampling strategy and choices."

Whether using large-N or small-N methods, humanitarian evaluators should explain their sampling strategy and choices. They should also explain any potential biases or other limitations inherent in their choices.

Yet what is considered a good enough sampling methodology? Are there certain programmes that lend themselves better to evaluation than others? For example, if the intervention uses mobile money and thus your program population all have cell phones is that easier to draw a random sample from? Do these more easily identifiable and traceable population or users lend themselves to a more rigorous sampling frame and what lessons can we draw from that? How few respondents can we live with to make assertions about the extended community? Are there better ways to create assurances that the information we have represents the views of the people we intend to assist? What other examples can we learn from?

Representative sampling is absolutely essential for quantitative methods where we want to generalise from the sample to the whole population. Representative sampling is also called probability sampling because in a representative sample, each person or family in the population

has an equal chance of being selected for the sample.

The ACAPS paper distinguishes between the indicators of quantitative and qualitative research as being: Internal validity/accuracy, External validity/generalisability, Reliability/consistency/precision, and Objectivity for quantitative research against Credibility, transferability, Dependability, and Confirmability, for qualitative research (ACAPS, 2013).

As noted, Daniel (2013) provides a thorough introduction to sampling. A good simple and basic resource for sampling is the *Practical guide to sampling* produced by the UK's National Audit Office (2000).

Goertz and Mahoney (2012; 2006) make the point out that qualitative and quantitative approaches are based around two different cultures, which although internally coherent, are marked by different values, beliefs, and norms. They also note that misunderstanding between quantitative and qualitative researchers "is enhanced by the fact that the labels quantitative and qualitative do a poor job capturing the real differences between the traditions. Quantitative analysis inherently involves the use of numbers, but all statistical analyses also rely heavily on words for interpretation. Qualitative studies quite frequently employ numerical data; many qualitative techniques in fact require quantitative information." They suggest that "statistics versus logic, effect estimation versus outcome explanation, or population-oriented versus case-oriented approaches" would be a better way of distinguishing the two approaches. One increasingly sees the terms "large-N" and "small-N" applied to the two approaches (Mahoney and Goertz, 2006).

"Qualitative and quantitative approaches are based on two different cultures."

Only large-N methods can answer questions about what percentage of a population got some particular benefit or saw a particular outcome. Only small-N methods can offer good explanations of why this happened. It could be argued that large-N methods with factor analysis can offer explanations, but only for those explanatory factors which are explicitly included in the study. This is why the best humanitarian evaluations use a mixture of both large-N and small-N methods to indicate both what happened and why.

References and further reading

ACAPS (2013) How sure are you? Judging quality and usability of data collected during rapid needs assessments (Technical Brief). Geneva: Assessment Capacities Project.

Daniel, J. (2012) Sampling Essentials: Practical guidelines for making sampling choices. Washington, DC: Sage Publications.

Longhurst, R. (2013) Implementing development evaluations under severe resource constraints. CDI Practice Paper. Brighton: IDS.

National Audit Office (2000) A practical guide to sampling. London: National Audit Office.

Puri, J. (2013) Experimental methods for impact evaluation. Presentation at ALNAP's 'Skills-building day for evaluators' on 14 March 2013 in Washington, DC.

Pritchett, L. and J. Sandefur (2013) Context Matters for Size: Why external validity claims and development practice don't mix. Center for Global Development.

Robert Wood Johnson Foundation. Qualitative research guidelines project. Website.

White, H. and D. Philips (2012) Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework. 3IE.

Department for International Development (2012) Broadening the range of designs and methods for impact evaluations. London: DFID.

This Method Note is the result of a discussion on 'Representative sampling in humanitarian evaluation', which took place among members of the ALNAP Humanitarian Evaluation Community of Practice (CoP) between December 2013 and January 2014.

ALNAP's Humanitarian Evaluation CoP provides a space to:
→ Discuss issues related to humanitarian evaluation
→ Share resources and events
→ Consult peers

ACCESS THE CoP HERE:
https://partnerplatform.org/humanitarian-evaluation

ALNAP
Strengthening humanitarian action through evaluation and learning