ALNAP

ALNAP Discussion Series

Improving the quality of EHA evidence

Summary Method Note

May 2017

Strengthening the quality of evidence in humanitarian evaluations

By Ian Christoplos

With Paul Knox-Clarke, John Cosgrave,
Francesca Bonino and Jessica Alexander

ALNAP is a unique system-wide network dedicated to improving humanitarian performance through increased learning and accountability. www.alnap.org

About the authors

Ian Christoplos is an independent consultant

Paul Knox-Clarke is the Head of Research at ALNAP

John Cosgrave is an independent consultant

Francesca Bonino is a Senior Evaluation Officer at UNHCR and a former Research Fellow at ALNAP

Jessica Alexander is an independent consultant

Suggested citation

Christoplos, I. with Knox-Clarke, P. Cosgrave, J. Bonino, F. and Alexander, J. (2017) *Strengthening the quality of evidence in humanitarian evaluations*. ALNAP Method Note. London: ALNAP/ODI.

Communications management by Alex Glynn and Maria Gilli

Design and typesetting by Chloé Sanguinetti

Copyediting by Nikki Lee

Contents

Introduction

High-quality, reliable evidence is central to humanitarian action. Evidence should inform programme choice and design, policy decisions, and the strategic direction of humanitarian responses and humanitarian organisations.

The **28th ALNAP Annual Meeting** in 2013 discussed the issues of generating and using evidence and knowledge in the humanitarian sector.[1] The meeting delved into questions such as: what do we mean by 'evidence'? 'Whose' evidence counts as valid and credible?

Evidence quality was a key issue tackled at the meeting. What do we mean by 'good quality evidence'? What makes evidence from across the humanitarian programme cycle credible and of sufficient quality so that different groups of intended users can trust it to inform decision-making and contribute to learning and programme improvement? What are the 'qualities' that are most important?

This paper builds on many of the concepts and ideas discussed in the **ALNAP Study on the quality and use of evidence in humanitarian action** (Knox-Clarke and Darcy, 2014), specifically focusing on evidence generated through evaluations. We draw on four Evaluation of Humanitarian Action (EHA) method notes that explore the evidential challenges confronting those who commission and carry out evaluation. Each method note addresses a different challenge:

1. Improving accuracy in humanitarian evaluations (including dealing with bias) [**EHA method note 1**]
2. Gathering representative data for evaluation purposes [**EHA method note 2**]
3. Ensuring quality of evidence generated through participatory evaluation processes in humanitarian contexts [**EHA method note 3**]
4. Addressing causation in humanitarian evaluation [**EHA method note 4**][2]

We assume readers will have some basic knowledge and exposure to evaluation concepts and vocabulary as presented, for instance, in the **ALNAP Evaluation of Humanitarian Action Guide** (ALNAP, 2016).

This paper is intended to be used by evaluators and evaluation managers, as well as by users of evaluations who need to judge and work to improve the quality of evaluation evidence. In particular, it will be used by those planning evaluations, as well as determining if and how to apply evaluation findings and recommendations in humanitarian programmes.

1. The ALNAP criteria on quality of evidence

In ALNAP's research, 'evidence' is defined as information that supports or challenges a given proposition. In the ALNAP Study **Insufficient Evidence? The quality and use of evidence in humanitarian action** (Knox-Clarke and Darcy, 2014), it is suggested that evidence in humanitarian action is important for three fundamental reasons:

- First, to ensure the quality of the response – for instance, gathering and using timely and credible evidence on the nature of needs and responses in crisis affected populations can enable responses to ensure effective coverage, meet priority needs and engage with existing capacities. In the longer term, evidence can inform learning for future programming.
- Secondly, evidence of humanitarian performance and results can feed into efforts to strengthen accountability to donors and partners – as well as to the affected populations.
- Finally, the process of gathering, interpreting and debating evidence of humanitarian performance and results can help humanitarian practitioners within their own agencies, partnerships and programmes, as well as within the so-called humanitarian system, to learn about and better understand their work.

Table 1: The difference between data, information and evidence

Data	Raw, unorganised facts
Information	Data that has been processed to show patterns and give meaning
Evidence	Information that relates to a specific proposition, and which can be used to support or challenge that proposition

Below is the overview and brief description of the criteria for good quality evidence proposed in Insufficient Evidence? (Knox-Clarke and Darcy, 2014). These criteria are generic, and are intended to apply to evidence collected using a range of methods.

Table 2: The six ALNAP criteria on the quality of evidence

Section	Criteria	Page	Definition
3	**Accuracy**	10	Whether the evidence is a good reflection of the real situation, and is a 'true' record of the thing being measured.
4	**Representativeness**	14	The degree to which the evidence accurately represents the condition of the whole population, the larger group of interest, the main stakeholders in the intervention, or the diversity that exists in the population.
5	**Relevance**	17	The degree to which a piece of information relates to the proposition that it is intended to prove or disprove.
6	**Generalisability**	20	The degree to which evidence from a specific situation can be generalised beyond a specific response to other situations (particularly important where evidence from one situation is used to create policies applicable to other situations).
7	**Attribution**	22	The extent to which the analysis demonstrates a clear causal linkage between two conditions or events.
8	**Clarity around contexts and methods**	27	The degree to which it is clear why, how, and for whom evidence has been collected.

As indicated by the sixth criteria, quality must be underpinned by commitments to transparency. Moreover, information can only be accepted as evidence if the methods used to gather and analyse it, and any limitations in the collection of evidence, are made explicit. Quality assurance (discussed in Section 2) is essential to ensure that evaluation is characterised by transparent and critically reflective processes.

All aspects of humanitarian action, whether it is early-warning, assessment, programme choice, implementation, monitoring, or evaluation, needs to be based on robust evidence rather than merely reflecting the prevailing narratives, biases or preconceptions. Indeed, evidence is essential in stimulating the critical reflection required to challenge the established narratives, biases and preconceptions that often steer humanitarian practice today. To this end, based on consultation and validation with ALNAP Network Members and desk-based research on evidence literature, Knox-Clarke and Darcy (2014: 67-69) suggest five principles to guide humanitarian practitioners' efforts in improving the quality of evidence generated at the different stages of the programme cycle. These five principles are:

1. **Using more robust methodologies for analysis and collection**: using tried and tested approaches from the social sciences for qualitative work, and continuing to explore the possibilities for quantitative and mixed methods approaches.

2. **Proportionate investment:** ensuring that investments in evidence match the importance of the questions addressed.

3. **Increased collaboration:** working together to identify key questions; decreasing unnecessary duplication and sharing and challenging results.

4. **Thinking of the longer term:** collecting consistent data and tracing trajectories sets over time.
5. **Including the knowledge of people affected by crises:** particularly by answering the questions that they need answered.

There are different entry points to putting these principles in practice. These include:

- **Exploring partnership with research institutions** to **improve methodologies** and evaluation designs; be **transparent** and **explicit** about the practical implications of decisions for mobilising researcher engagement in various considerations over quality.
- **Collaborating** to identify gaps, share results and critique findings; **reach out** across different evaluation communities to overcome disciplinary **silos**.
- **Building evidence over time** – evaluations should not be standalone events but should **draw on rigorous and appropriate baseline and monitoring data**, and contribute to **broader tracking of results** beyond individual projects.
- **Addressing systematic exclusion of the voices of crisis affected people** both as a basis for ensuring their **rights to participate** in judging humanitarian programming and to **overcome tendencies to accept the often unquestioned agency narratives** that drive programming.

Evidence and methods – quantitative and qualitative

There is no simple methodological solution that can guarantee the quality of evidence. Indeed, quality will be closely linked to the appropriateness of a method used in the evaluation. Any method can be used in a more or less rigorous manner, and the level of rigour will be related to available resources (time, human resources, etc.) linked to the scope and ambitions of the evaluation. In many cases, quality is strongly associated with having enough lead time to access and read documents, arrange interviews and have a dialogue with stakeholders – not least in order to better understand the 'qualities' of evidence that they find useful and credible.

Much of the debate around the quality of evidence concerns the relative merits of quantitative and qualitative methods, and the opportunities to combine different methods in a single evaluation. Goertz and Mahoney (2012: 2006) 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. 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' (ibid.). 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 (ibid.).

The size of the sample, referred to as 'large-N' and 'small-N', is central to the debate around the two approaches (Mahoney and Goertz, 2006). Only large-N methods can answer questions about what percentage of an overall population received particular benefit or saw a particular outcome. While 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. Section 7 looks further at the different ways that evidence can be used to support or rule out assumptions about causality.

2. Ensuring evidence quality as part of quality assurance

Offices commissioning evaluations, firms undertaking evaluations and others engaged in the evaluative process have a responsibility to exert due diligence over evaluative work and support evaluation teams to collect good quality evidence. This requires careful management of levers at different points in an evaluation process. A range of entry points exist for those responsible for quality assurance to influence the quality of evidence generated from evaluation work.

Tips for evaluation departments and commissioners:

1. Demand that evaluation teams provide detailed methodologies as a solid basis for discussions on efforts to ensure the quality of the evidence. Open and fruitful discussions around methodologies during inception phases can enable the same constructive dialogue about the quality of evidence throughout the evaluation process.

2. Promote evaluation policies that encourage dialogue between evaluators and the intended users to better understand what constitutes credible and useful evidence.

3. Set a modest and realistic number of relevant evaluation questions, reflecting the time and resources available. Almost inevitably, the quality of evidence will be in reverse proportion to the number and complexity of the evaluation questions. Where many stakeholders are involved, the number of evaluation questions can mushroom, leading to a 'thinning' of the evidence base for each question. Evaluation commissioners may need to play a gatekeeper role vis-à-vis their colleagues to maintain realistic expectations.

Tips for firms undertaking evaluations, commissioners (where relevant) and team leaders:

1. Set the evaluation team selection criteria to encourage capacity for gathering and analysing robust evidence, and to develop the links needed to influence programme teams and policy-makers.

2. Allow time for planning, inception phases and a thorough review of the evaluations' inception reports. This should also provide space to reflect on how the team can best conduct a robust and credible evaluation. The inception process should factor in the strength of evidence needed, and the political context in which the evaluation will be used, as well as the need for dialogue with programme teams and policy-makers.

3. Insist that evaluation teams are 'not shy' (especially during inception phases) in being clear about constraints and limitations to achieving often ambitious evaluation objectives, while maintaining a high standard for the quality of evidence.

4. Allow time for debriefing/verification meetings in the field to check facts and identify major gaps or misconceptions.

5. Provide resources and time for independent quality assurance functions, including a thorough review of the inception and evaluation reports. The inception report should emphasise analysis of methods, sampling procedures and selected indicators and data sources. The evaluation report should focus specifically on the quality of the evidence and analysis offered. In particular, this involves checking the evidence base sufficiently supports the conclusions and recommendations presented in the report.
6. Plan for dissemination events to support the take up and use of the evaluation.

These actions can be seen as necessary ingredients to improve the 'supply' of higher quality evidence in EHA. However, they are not sufficient on their own to encourage and support demand for high-quality evaluative evidence to improve programming and results. Quality assurance, therefore, should happen throughout the evaluation process to ensure that all opportunities are exploited to engage the intended users' feedback on the 'qualities' of evidence that encourage use within their organisations. Some humility is important too, as organisations learn (or fail to learn) from evidence in evaluations in a variety of ways (Hallam and Bonino 2013). This is an important aspect of the search for 'methodological pluralism and appropriateness' to adapt to different users' needs (Patton, 2014). Moreover, evidence quality cannot be separated from concerns over considering what forms of evidence are meaningful, credible and useful for different actors within different processes happening in the commissioning organisation, and the frames of reference of different users.

3. Accuracy

Whether the evidence is a good reflection of the real situation, and is a 'true' record of the thing being measured.

One of the main tasks for evaluators is ensuring that the data gathered is as accurate as possible. But putting a finger on accuracy is difficult, as evaluators must gather 'accurate' data on both objective 'facts' and subjective 'perceptions'. Regarding the latter, an accurate measurement of people's hopes and fears in responding to a humanitarian emergency may be at least as important as the 'facts' of loss of life and other factors normally associated with 'accuracy'. The ALNAP Study Insufficient Evidence? defines accuracy as:

> *'Whether the evidence is a good reflection of the real situation, and is a 'true' record of the thing being measured. Anthropometric measurements that have been conducted correctly, and key informant statements that are a true account of what the informants believe, both pass the test of accuracy. Accuracy can be compromised by the informant or by the observer/researcher, through conscious falsification, unconscious bias, misunderstanding, or the incorrect use of measuring instruments.' (Knox-Clarke and Darcy, 2014: 15)*

To qualify as accurate, evidence should be a good reflection of the real situation, whether that refers to facts on the ground or informant beliefs, or indeed (as will be discussed below) triangulation and comparison of the two. Evaluators may encounter problems with accuracy when conducting interviews, or when using secondary data.[3]

Bias

With respect to interviews, it is important to understand why respondents may fail to tell the truth, either consciously or subconsciously. If the line of questioning is threatening to them – i.e. it exposes limited skills or knowledge or highlights socially unacceptable behaviour – a risk of what is called 'desirability bias' exists. For example, a mother may not want to tell the truth about how often her children have been fed if she believes that the truth would reflect poorly on her. Another (related) source of bias is the respondent's idea of how his/her answer may affect their environment such as a decrease in aid or a change in programme. For example, respondents may exaggerate claims about their conditions and problems if they believe it can further their wellbeing. This may be a factor with both affected populations or with agency staff eager to retain their jobs. The presence of members of the family or community (or if higher level authorities or other powerful actors are likely to be informed about responses) may also influence the accuracy of how respondents answer questions. Rather than giving a truthful reply, the respondent may be tempted to answer in a way that gives him/her credibility or respectability in the eyes of onlookers. Other reasons for receiving skewed information could stem from faulty memory, misunderstanding of the question or purpose of the interview, or a desire to give answers that the respondent thinks the interviewer wants to hear, or 'courtesy bias'.

Even analysis of secondary data is prone to inaccuracy due to similar factors. For example, assessment data from a study in Ethiopia found that a government needs assessment had an economic slant, and was more of a 'wish list' than an accurate assessment of needs. In response, Clusters and local NGOs played a vital part in building confidence in government data and its accuracy (Darcy et al., 2013). Moreover, in other cases such as in nutritional surveys, data that can be interpreted as revealing failures (by humanitarian agencies or government) might even be suppressed, and the range of secondary data limited to presenting only the positive trends.

Accuracy is inevitably affected by the frames of reference of the person or document providing the information. When presenting the evidence, the providers and users of the information may have a different set metrics by which to measure the factors that are being judged. An accurate measure of 'sufficient' food, shelter, security, etc. is often related to the 'indicators' perceived to be relevant by the informant. Aid recipients are often positive towards receiving what they expect the agency wants to provide, and may avoid or be unaware that suggesting different forms of support is possible. This potential bias can be recognised with sensitivity regarding the person being interviewed, including awareness of the geographical area to which they are referring, and making sure that poor, marginalised ethnic groups, disabled people and women are well represented, at a minimum as subjects of the inquiry.

Biases and preconceptions on the part of evaluators can have profound effects on the quality of evidence. These biases may be related to gender, ethnic or other 'lenses' that skew how they perceive the situation under analysis. Gender and ethnic diversity within evaluation teams are important to counteract these tendencies, but are not panaceas. Explicit methods to look at discrimination are the most important means of overcoming these internal biases.

Some examples of these methods come from a case study in Kosovo (see Westley and Mikhalev, 2002) where the team identified significant bias when interpreting the results due to ethnic prejudice between Serbian and Albanian team members. Upon reflection, the evaluation team recommended the following for promoting better shared understanding and minimising bias:

- Showing **videotapes** of the sites and assessment work in the different ethnic areas to each team.
- Holding carefully arranged **meetings** between team members if they were willing and interested.
- Making an effort to **recruit other minority groups** into the assessment.
- **Recruiting more individuals** who were comfortable working across the ethnic divide and provide them with the necessary security and protection.

Other evaluation team biases may stem from (conscious or unconscious) tendencies to judge interventions based on the extent to which they simply 'tick the boxes'. These often relate to client expectations and resemble models of 'best practice', also referred to as 'isomorphic mimicry' (Andrews et al., 2012), rather than relying on the extent to which overall objectives are met and organisations are actually performing effectively. Failure to look critically at the relevance of overall approaches in the intervention for fear of criticising the client or questioning recognised orthodoxy around 'best practice' may mean that indicators are selected that are biased towards judging whether programme participants and those using services are 'doing as they are told to do'. This enables negative self-censuring analyses of the more sensitive issues.

Triangulation

Most humanitarian evaluations rely on some form of triangulation as the primary means to enhance accuracy. Triangulation has been defined as 'deliberate attempts to confirm, elaborate, and disconfirm facts and interpretations through multiple data sources, multiple methods of data collection, repeated data collection over time, and achieving multiple perspective through involving/consulting multiple researchers' (Bamberger et al., 2012: 137). According to **BetterEvaluation**, triangulation is a way to validate data by testing:

> *'Consistency of findings obtained through different instruments and increases the chance to control, or at least assess, some of the threats or multiple causes influencing our results. Triangulation is not just about validation but about deepening and widening one's understanding … It is an attempt to map out, or explain more fully, the richness and complexity of human behaviour by studying it from more than one standpoint.'*[4]

An assessment of the situation in Aleppo in Syria provides an example of rigorous triangulation. Enumerators were asked to assess the reliability of key informants, and triangulate informant responses with each other, with secondary sources and with their own observation. Where discrepancies occurred that could not be resolved, the information from the informant was not used (Assessment Working Group for Northern Syria, 2012).

Although most evaluations mention triangulation as central to ensuring accuracy, they rarely describe how triangulation is done in practice; what this means for the findings; how much weight is given to different data sources; and whether it is possible to draw reasonable conclusions based on them. Transparency is key to determining the quality of evidence and many evaluators are lax about showing the details of how they intend to apply triangulation, even if they claim that triangulation is critical to their analysis in the methods section of their reports. Poor triangulation methods can have considerable implications for how evidence is interpreted. While triangulation can be used to screen the quality of data, as in the Aleppo example above. It may also be used to bring out the range of perceptions of a given phenomenon, as noted in the following:

> *'Contradictions and inconsistencies are often discarded or categorised as untrue, when their analysis could be of great value and represent variability in the observed setting. When using qualitative research to study a complex phenomenon, social researchers suggest that convergence of data should not be expected when performing triangulation. Instead of a tool for validation, the triangulation process should offer different perspectives and give access to different versions of the phenomenon under scrutiny.' (ACAPS, 2013: 17)*

These alternative perceptions may be seen as unwelcome dissonance by evaluators under pressure to present unequivocal answers about 'results' and 'what works'. However, divergent perceptions of what constitutes appropriate and valued results may be essential for recognising where a project might be excluding certain groups or leading to a combination of positive and negative outcomes.

In some cases, a particular approach to triangulation is checking the accuracy of answers by measuring them against a 'constant' or another known value. Depending on the nature of the evaluation, and the availability of other data, evaluators might correlate information from interviews against known values such as prices, seasonal trends, nutritional status or morbidity rates (while recognising that inconsistencies may, as likely, be a result of poor data as of inaccuracy in interviews). Sometimes, areas of inquiry can be verified by evaluators themselves. For example, interviewers can walk the distance that respondents have to go to get water in order to verify the distance reported. Measuring against constant or known values in this way can be useful in considering the likely margin of error in data, and its significance for the conclusions being drawn or the calculations being made.

Accuracy in secondary data

The secondary data that is most relevant for humanitarian evaluations can often prove to be inaccurate. This may be because there were methodological errors in the collection of the data, limited samples, biases (outlined above) related to selective or misleading release of data for political purposes, or simply because the data is out of date (often grossly out of date in long-running chronic conflicts). The secondary data collected by the agencies being evaluated often largely consists of baseline and monitoring data that may have focused on the wrong indicators, or can be a result of haphazard data collection or weak commitments to monitoring. Again, triangulation can be helpful here: one piece of secondary data can be triangulated against others, or against known values.

Equally important, however, are the methodological statements made in the secondary data reports themselves. Any data set or report that is presented without a clear statement of the methodology used for collection and analysis should be treated as inherently poor in quality. Where a method is given, it may be important to make a judgement of the quality of the data based on the apparent rigour of the method and how it was applied. For data that rely on statistical tests, this is fairly easy to do (as long as a member of the evaluation team has the requisite skills and knowledge in this area). For qualitative data, it may be harder to determine the quality of the research from a method statement. There is no single, accepted 'hierarchy' or system of grading for qualitative work.

4. Representativeness

The degree to which the evidence accurately represents the condition of the whole population, the larger group of interest, the main stakeholders in the intervention, or the diversity that exists.

Quality evidence builds on data that is representative of a given population or set of informants at a certain time, and which can be used in comparison with other snapshots in time. Representativeness refers to the degree to which the evidence (often from a sub-set of the population) accurately represents the condition of different groups within the overall population. This has two aspects. First is to judge whether this sub-set is representative of the majority of the population. Second, is to assess whether certain groups are excluded or benefit unduly from a given intervention. Quality in relation to representativeness is largely related to controlling potential bias, which may occur because of under-coverage of some groups of informants due to lack of access or insufficient time and resources. When undertaking surveys, an example of bias would be an underestimation of income levels because those working longer hours in the sampled population have a lower response rate (ACAPS, 2013). Pastoral groups too may be under-represented if their settlement patterns are not taken into consideration in designing the methods. While people with disabilities may be hidden from sight. Certainly challenges in obtaining representativeness can be overcome – for example, enumerators can return to homes where someone is not there instead of moving to the next house where circumstances may be different; or data can be collected at different times of the day and during weekends to maximise reach to a representative group.

Selecting an approach to sampling

Evaluators can rarely interview all relevant stakeholders or assess the whole of the affected population or all cases assisted. Data is normally gathered from a sample as a means of drawing conclusions about the perspectives of a large range of important stakeholders, including crisis affected populations targeted by an intervention, or even the population as a whole. If sampling is not conducted effectively, evaluations are likely to over- or underestimate the extent of the issue being described, which threatens the quality of the evidence presented. If the evaluation uses statistically-based methods, it is important that sampling is properly done to draw statistically valid inferences from the expensive data collected.

When collecting data from authorities, agencies or service providers, it may be possible to interview a large proportion of the 'stakeholders' concerned with a given factor in the evaluation in a way that their 'representativeness' might not be problematic. When gathering data from affected populations, intended aid recipients or (especially) those marginalised or excluded by the intervention being evaluated, representativeness becomes more difficult. For this reason, sampling methods that can yield the highest quality of evidence possible with the resources available need to be chosen.

In essence, there are two overall types of samples: probabilistic and purposive. Probabilistic sampling uses various methods based on random selection to obtain a representative sample among the entire population, or among a large affected population. This method allows the evaluator to know the probability of the sample is representative of the population as a whole, and so allows for statistically reliable generalisation. Purposive sampling – consciously choosing particular groups or individuals, rather than selecting them at random from the population – is generally used when the evaluation focuses on a particular set of affected people, or sets of 'stakeholders', in order to analyse issues of particular interest. It is based on the idea that the people or groups

sampled will be particularly knowledgeable about the issue of interest – for example, the extent to which they were able to access and benefit from the intervention, or whether it was relevant to their needs.

Ensuring quality in probabilistic sampling

Probabilistic sampling methods include random, stratified and cluster samples. A defining characteristic is that the probability of inclusion of every member can be determined (in random sampling, it is equal for every member). Probabilistic sampling is generally used for quantitative studies and chosen where a large sample size is needed, where the characteristics of whole population are known, and where the factor under analysis is broadly distributed in population. These methods are also important where broad representation is required, resources are available, and access is largely unrestricted. Needless to say, this is not the case in most humanitarian evaluations. These methods are useful to respond to evaluation questions asking 'how many?' and 'what percentage of population?'

The **Evaluation of Humanitarian Action Guide** (see Section 12, for discussion on sampling) states that unless a sample is randomly selected, it is not possible to make valid generalisations to the whole population because the sample is not representative. 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 from which that random sample can be drawn.

In many situations evaluators may not, for a variety of reasons, be able to ensure representativeness through the use of a random sample, and may instead aim to create an illustrative picture of the situation of a population. Alternatively, the evaluation may not require a sample that represents the whole population but rather one that reflects the experiences and knowledge of a particular group (such as young women, or field staff). Under these circumstances, purposive sampling methods are more appropriate.

Ensuring quality with qualitative approaches

Even when using qualitative methods, careful decisions need to be made about where data is to be collected and who to interview. Purposive sampling for qualitative analyses usually starts with a form of stakeholder analysis anchored in contextual review and in-depth consideration of who has a 'stake' in (and who may have been excluded from) the intervention. There should also be a particular focus on 'stakes' impinging on the issues to be addressed in the evaluation. Evaluations that primarily involve collection of data from key informants and focus groups need to be based on explicit criteria about why a given informant is 'key', and why the particular focus is of relevance for the evaluation.

It cannot be stated with certainty that the people or households in a purposive sample are representative of the entire population of interest (as noted above, this is the role of probabilistic sampling). However, a purposive sample can be used to test the legitimacy of the theory of change or intervention logic of a programme in relation to specific groups. By purposively and rigorously analysing how different sub-groups are influenced by and are influencing implementation of an intervention, it is possible to draw general conclusions about the extent to which the theories behind the intervention are valid. Effectively, the sample cannot represent everyone in the population i.e. it cannot be said that everyone in the population is 'like' the people in the sample. But comparing the extent to which the logic is sound in relation to different sub-groups, a well-structured purposive

sample often provides clearer insights into 'why' an intervention has been successful or not, as compared to a probabilistic sample that tends to emphasise yes/no answers. The difference in approach has been described by Yin (2010) as the difference between 'statistical generalisation' and 'analytical generalisation'.

With purposive sampling, interviewees or cases are selected to obtain rich sources of information about the situation or views of one set of 'stakeholders' or population group – such as service recipients/clients, community leaders, or people who have extensive experience with the population being examined (such as mothers and teachers if studying behaviour of children). Purposive sampling is primarily used for qualitative analyses, as the selection of informants is intentionally biased and reflects those with a particular knowledge or interest in the topic being analysed. Even if the evaluation selects a purposive sample in terms of primarily interviewing agency staff or local authorities, it is important that evaluation readers are made aware of the sampling procedures so they know how many people were interviewed and how they were selected. Gender, education levels or other variables may be important.

Effective purposive sampling relies on the principle of 'saturation'. This means that an understanding of the situation has been achieved when all perceptions have been uncovered, and when collecting new information does not add more to the understanding of the situation. The principle of saturation also implies that sampling, information collection, analysis and resampling are an iterative process. To establish when saturation has been achieved involves analysing the information collected during the actual data collection phase. It also requires alertness to new and divergent information, as this may mean that another subgroup has been encountered with a different experience of the programme being evaluated. Where this is the case, more interviewees may need to be added to the sample.

Qualitative approaches are typically characterised by small-N. For these cases, snowball sampling is a good method, particularly in finding key informant interviews. This method uses chain referral sampling, where each interviewee is asked to suggest the names of other interviewees who can speak authoritatively on the topic. This can lead to dozens of interviews as the chain continues until no new names emerge. In these cases, the evaluators purposefully seek out people who can provide the most amount of information. If the evaluation wants to know what factors contributed to success in a microfinance project, for example, this may bias the sample by focusing interviews on the most successful users of credit and asking them what led to their success. This means that the evaluation cannot generalise to the whole population, or uncover the factors that lead to some being excluded from the services. It may, however, provide data on success factors. This is an example of the challenges that many evaluations face in ensuring that the voices of 'non-stakeholders', for whom the intervention has not yielded benefits, are not overlooked.

Convenience sampling

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 the evaluators happen to be present – and which may not allow credible inference about the population. Security issues or overambitious evaluation scope in relation to time and human resources for the evaluation, may sometimes necessitate a convenience sample. It is important to acknowledge the limitations inherent to this approach.

5. Relevance

The degree to which a piece of information relates to the proposition that it is intended to prove or disprove.

Relevance may seem straightforward, but the criterion becomes complicated in evaluation practice. The relevance of a particular piece of evidence may only become apparent when the evidence is analysed in relation to the evaluation questions. It is only then that the value of the evidence becomes apparent.

The relevance of a piece of evidence depends on how it relates to – and ultimately interrogates – the assumptions behind the intervention. Nearly all humanitarian and development programmes are based on (explicit or implicit) theories of change about how desired activities and outputs are expected to contribute to intended outcomes and impacts. Both monitoring data and choice of indicators for the intervention generally tend to mirror these underlying assumptions and ask whether actors are 'doing things right' in relation to the intervention logic, or even a policy.

Evaluators should go a step further and take a critical stance in relation to the theory of change. They can do so in selecting and collecting data that unpacks the relevance of these assumptions, and take an independent view on the relevance of the programme's own data. They should not assume that data showing that activities have been implemented will lead to intended impacts since the intervention may not be 'doing the right thing'. For example, many agricultural rehabilitation programmes claim to contribute to food security. Data is often collected to assess whether the distribution of high yielding varieties of seeds has led to an increase in grain production. However, this data may not be relevant for judging whether the seeds have led to enhanced food security if markets were not considered, or if the changes in the production system led to indebtedness when smallholders became dependent on purchasing fertiliser, or if growth of markets led to land grabbing. Relevant evidence should reflect the big picture.

Efforts to address the relevance of evidence, therefore, often leads the evaluation team back to analysis of the initial results framework for the intervention. The intervention results framework may be presented as the basis for selecting indicators to evaluate against, but these indicators may reflect initial assumptions about the intervention regarding the conflict, the disaster, or the capacities, needs and aspirations of the populations or organisations engaged in the evaluation, that are proven to be misguided during implementation. Even when and where the initial indicators in the results framework have been found problematic early on, the monitoring system for the intervention may nonetheless be 'stuck' in collection of irrelevant data that the evaluation cannot use.[5] For example, it may have been recognised that food aid is not reaching the most vulnerable populations, but a monitoring system that uses quantities of food distributed as a proxy for improved nutrition may remain in place.

Another aspect of relevance involves questions regarding whether the standards being used are relevant for achieving overall policy objectives in the given context. This may relate to humanitarian standards, or even lead to questions about whether the focus of the intervention was actually relevant to needs – e.g. did the intervention provide services such as food or water, when relevance in the intervention context actually suggests that there was a greater need for protecting the overall policy goal of alleviating human suffering? As such, indicators that give evaluation teams the latitude to ask these 'big questions' are essential for ensuring that relevance is appropriately assessed.

Indeed, these challenges may be turned to an advantage if the evaluators are given the leeway to collect data that effectively questions the underlying theory of change of the intervention. Ideally, theory-based evaluation methods (see Section 7) apply indicators that critically assess whether the intervention has been 'doing the right thing'. The Grand Bargain agreed upon at the World Humanitarian Summit reflects broad questioning about whether the humanitarian sector is 'doing the right thing'. Evaluations of relevance in relation to the Grand Bargain can provide important information about these fundamental questions, and may even provide evidence on whether the Grand Bargain is the 'right alternative'.

Whose indicators?

Answering questions about 'doing the right thing' requires selecting indicators that reflect other perspectives on relevance than those described in project documents.

> *'While 'traditional M and E systems tend to over-emphasise "our indicators" not "their indicators"' (Catley et al., n.d.: 21), the participatory approach uses impact indicators identified by the intended beneficiaries. Similarly, the People First Impact Method (PFIM) asks communities to identify the most important changes in their lives, and the causes of these changes (for example, see O'Hagen and McCarthy, 2012). A similar approach was used by the Food and Agriculture Organization (FAO) in Somalia where the impact evaluation relied on an iterative and narrative based approach: aid recipients discussed their experiences of the programme, and the impacts it had, in semi-structured interviews that were enhanced by a variety of participatory tools. Emerging themes were then fed back to community groups for further discussion and verification (Tessitore, 2013).' (Knox-Clark and Darcy, 2014: 43)*

In a recent volume discussing progress and application of participatory methods, Jeremy Holland points out that 'in contrast to the individualised observations and discussions in much top-down investigation, participatory research also focuses on public and collective reflection and action' (Holland, 2013: 2). He and other researchers have argued that participatory methods can both generate accurate and generalisable statistics in a timely, efficient and effective way, while also empowering local people by using methods that facilitate local control of data generation and analysis. The assumption is that the stakeholders' involvement helps ensure that the evaluation addresses appropriate issues and (due to enhanced relevance to 'their indicators') gives them a sense of ownership over the results.

Stakeholder involvement in ensuring that indicators are relevant for the local context and the specific intervention has been shown to also increase the chances of evaluation results being seen as relevant by programme decision-makers and implementers (for example, see Patton, 2007). This may even lead to the evaluation being learning experience for the programme stakeholders, increasing their understanding of programme strategy, and contributing to improved communication between programme actors working at different levels of programme implementation (Aubel, 1999).

The literature considers a truly participatory evaluation to be one where the affected community is involved in ensuring a focus on relevance in all aspects of the evaluation: planning and design, gathering and analysing data, identifying findings and formulating recommendations, disseminating results and preparing a plan to improve programmes.

Guijt and Gaventa highlight four principles that are at the core of participatory approaches to evaluation (1998: 2):

- **'Participation'**: opening up the design of the evaluation process to include those most directly affected, and agreeing to analyse data together.
- The inclusiveness of participatory monitoring and evaluation requires **'negotiation'** to reach agreement about what will be monitored or evaluated, how and when data will be collected and analysed, what the data actually means, and how findings will be shared, and action taken.
- Leading to **'learning'**, which becomes the basis for improvement and corrective action.
- Since the number, role and skills of stakeholders, the external environment, and other factors change over time, **'flexibility'** is essential.

Participatory evaluation is not a panacea for achieving relevance

Despite these notable advantages in employing participatory approaches to promote relevance, there are some pitfalls. Those who are close to an intervention may have 'too much ownership' and, as described in Section 3 above, can lead to bias and inaccuracy if they are dedicated to defending the flow of resources flowing from the intervention.

Any intervention will have different levels of relevance for different sectors of the population. If representation is skewed, then the participants may also fail (either intentionally or unintentionally) to ensure that the interests of 'non-participants' are represented in the evaluation. This relates more generally to even non-participatory evaluations if the focus is on soliciting information about intended 'beneficiaries', when the needs the intervention is intended to assess may be greater among populations that have not benefited.

These considerations relate to the need to retain a focus on independence to ensure the broader issues of relevance are not overlooked, even while the evaluation strives to ensure that relevance among local stakeholders is adopted. Perhaps paradoxically, independence is also important for creating the distance to place the intervention within a broader context of analysis within the prevailing political economy of the conflict, the wider trends in economic development, demographic trajectories or climate change. These 'big picture' issues may be beyond the perspectives (or simply taken for granted) by local actors. Yet, these issues are essential for ensuring that a balance is maintained between local relevance and what can be 'relevantly' generalised to other contexts and interventions, as will be discussed in the following section.

6. Generalisability

The degree to which evidence from a specific situation can be generalised beyond a specific response to other situations.

Generalisability concerns the scope of the proposition that a piece of evidence supports or refutes. It involves extrapolating the relevance of findings and conclusions from one intervention, local context, conflict, disaster or policy to others. As such, qualities relating to generalisability have as much to do with the way that evidence is analysed as they are about the quality of evidence itself. Generalisability in evaluative practice can be looked at from two perspectives.

First, generalisability involves looking at similar processes, causal mechanisms or theories of change in different interventions or contexts. The commissioner or organisation may want to know if the model or method applied in the intervention being evaluated can be applied elsewhere (e.g. cash-based approaches or a particular system for training local staff). Here, the aim can be to generalise from one context to another context (Does 'what works' in South Sudan also work in the Democratic Republic of Congo?); or to generalise from one context to a large set of contexts (Does 'what works' in South Sudan also work in humanitarian responses in general?).

Second, thematic evaluations use meta-analysis of a variety of evaluations to look for patterns in relation to a theme (e.g. 'coping', 'recovery', etc.), where generalisations can then be made in relation to broad policy assumptions or narratives that frame a range of different interventions. Less focus is given to the generalisability of a given model or method. Instead these approaches look at a variety of models and methods to identify the factors (often beyond the intervention itself) that have influenced success or failure.

There is an active discussion across the social sciences about the methods and types of analysis that are most useful for generalising research findings. Case study evidence, for instance, involves different types of generalisability than, say, quantitative analysis, and therefore different measures of rigour. Recent work to improve the rigour of case study methods have suggested several techniques for widening the scope of case study findings, including pattern analysis (Yin, 2008; Blatter and Blume, 2008); and structured comparison (George and Bennett, 2005). Observers often question the value of case studies for generalisation as they may appear to emphasise features that are only relevant to the given case. For this reason, particular attention is needed when structuring evidence to contribute to more generalisable conclusions.

The extent to which case study findings can be confidently generalised depends on the application of thorough and consistent analyses of two dimensions. First, the evaluators should explicitly describe the concepts and theories behind the interventions so that 'patterns' in these concepts/dimensions can be identified and compared. Second, in addition to looking at the concepts framing the intervention, patterns in the actual empirical evidence within the set of cases being analysed need to be identified and compared. In the example of food security, the evaluation should compare patterns in how programming has been anchored in theories about how to enhance the basic determinants of food security, such as how increases in production or use of new technologies are expected to improve access, availability, utilisation and stability. Second, case study analyses should also identify patterns in the empirical evidence of how the programme has actually influenced markets, shifts in reliance on subsistence farming, and access to specific types of food among different target groups. Together, analysis of the theories behind the intervention and the empirical findings about outcomes in the field should provide a basis for generalising about different dimensions of food security outcomes. Generalisability

thus involves using evidence to describe patterns in both programme theories and in observed phenomena in different cases to judge where evidence points to factors that can be generalised (or not).

There are inevitable tensions between local relevance and generalisability, and evaluators may need to make difficult choices regarding how to prioritise the two different 'qualities'. Newer developments in case study approaches may offer a way for evaluators to 'have their cake and eat it'. These approaches can focus on local relevance while actively looking for ways to structure their analyses and presenting evidence that can enable their readers to recognise patterns informing how they think about their own contexts and programming.

Here again, triangulation may be important to provide support for generalising from very local cases. Mixed-method approaches can provide additional evidence on where there are similarities and differences in populations, factors related to the conflict or disaster, or service provision performance that relate to the policy, context or type of intervention with which to make comparisons. Furthermore, as will be discussed in Section 7, the quality of analysis of attribution and contribution provides evaluation users with the confidence they need to judge whether similar causal assumptions can be presumed to be valid in other situations.

Finally, it should be stressed that it will ultimately be up to the users of any evaluation to decide if and how to generalise the findings to other settings and applications.

7. Attribution

The extent to which the analysis demonstrates a clear causal linkage between two conditions or events.

Why is attribution important?

Attribution seeks to establish the extent to which particular actions have caused particular effects. In other words, it is about being able to describe and justify the links between causes and effects (Stern et al., 2011:1). As Knox-Clarke and Darcy note, 'It is not enough for an evaluation to depict a situation accurately; it also needs to show the relationship between a specific intervention, or series of interventions, and the situation described.' (2014: 40). The intention of logical frameworks and many other tools for designing results frameworks is to force programme designers to be transparent and explicit about the causal assumptions behind the intervention.

In evaluation, attribution is important because it is about seeking to establish what difference an intervention makes or how the intervention contributes to changes that are underway as a result of a range of factors. When evaluations answer causal questions, they need to judge what a programme had set out to do (i.e. the results framework) based on what it actually achieved, and how effectively. In order to establish causation, evaluations need to articulate the intervention's role (even if this is limited) in producing or contributing to certain results in terms of outcome or impact. If particular outcomes are observed, to what extent can the outcomes be verifiably attributed to the intervention? Put differently, addressing causation in evaluation is about documenting that a given result, change, or effect has been caused by the intervention and not by coincidence or by other concurrent factors at play in a given context (Davidson, 2005).

Programme managers, donors and decision-makers at all levels want to answer whether and to what extent their interventions lead to the desired results, outcomes and impacts. However, not all evaluations address, or are expected to answer direct questions about cause and effect. Evaluations may, for example, focus on cost efficiency or whether conditions have been created that are likely to lead to sustainability in the future even if it is too soon to make judgements about whether these elements have 'caused' sustainability. Many observers have noted that most evaluative work focuses on process issues, activities and outputs i.e. describing what happened during an intervention and how, rather than answering the question of whether the intervention brought about the intended results (outcomes and impacts) (Morra Imas and Rist, 2009: 228). Attribution is often important in various types of process evaluations as they (should) ask about the extent to which the process that is underway in the intervention stimulated, supported, or hindered the intervention. However, the ultimate impact on affected households may not be given priority attention in the evaluation as these impacts are seen to be phenomena that will arise in the future and/or relate to a range of concurrent factors that are not possible to assess. This points to the importance of ex-post evaluations to assess whether benefits reach crisis affected and/or marginalised populations. Yet regrettably, ex-post evaluations of humanitarian assistance programmes are relatively rare.

Attribution and contribution: two overlapping measures of causality

Evaluators often need to be able to document that a given result, change, or effect has actually been caused by the intervention and not by coincidence. Establishing this causal relationship (also called **causal inference**) is not straightforward. It may not be possible to isolate the results brought about by a given intervention in among a host of other factors at play, in a given context. Randomised control trials (RCTs) can provide strong evidence

about a single variable. However, they are less effective when dealing with causality in complex situations where factors such as violence, drought and state fragility are part of the intervention itself. This complex and rapidly changing mosaic of phenomena can profoundly influence the contribution of the intervention in addressing the humanitarian needs of different sub-groups of the population. This points to what is commonly referred to in evaluation as the **attribution problem**, and is the focus of an extensive body of literature in the evaluation field. Attribution requires establishing the causal implications of an intervention and/or the causation of a phenomenon that can be observed (Scriven, 2010: I).[6] However, establishing attribution does not equate with establishing 'sole attribution' – an intervention solely necessary and sufficient for a particular effect. Establishing causation is an issue of wider concern across the social sciences (Illari and Russo, 2014) and some have addressed this through different approaches, such as process tracing to identify causal mechanisms (Beach and Pedersen, 2013), or by exploring **contributory causation**.

Establishing causation is a particular challenge for evaluators across the many branches of evaluation practice – but within the humanitarian sector, there are even greater complications when addressing cause and effect questions. Given the rapidly changing and often unstable environments where humanitarian actors operate, establishing and tracing the contributions of modest programmes in contexts where affected populations experience multiple risks is inherently difficult. For example, when looking at reduced malnutrition, there could be a range of factors at play contributing to the results, such as: other concurrent health-related interventions (e.g. de-worming); improved/declining hygiene conditions; seasonal factors; and changes in household income. Insufficient analyses of the conflict, political processes and economic trends can lead evaluators to miss many of the major causal factors behind the apparent results of the intervention.

In most cases, evaluators must assume that there are numerous factors at play aside from the intervention that make it impossible to conclude that the intervention alone 'caused' a result. A more realistic expectation is that the intervention, together with other influencing factors (what Cartwright and Hardie [2012] and Mayne [2012] refer to as a 'causal package') has contributed to the outcome. Mayne thus suggests that these causes together, which are neither necessary nor sufficient, can be called **contributory causes** (Mayne, 2012: 2). From this 'contributory' perspective, appropriate evaluation questions would be:

- Was the causal package of the intervention plus its supporting factors sufficient to produce the intended result?
- Is it likely that the intervention has made a difference?
    - » Is it likely that the intervention was a contributory cause of the result?
    - » What role did the intervention play within the overall causal package?
- How and why has the intervention made a difference?
    - » How did the causal factors combine to bring about the result?
    - » How did the context affect this contribution and which mechanisms were at work?
- Has the intervention resulted in any unintended effects? (Mayne, 2012: 1-2)

BetterEvaluation notes that the essential value of contribution analysis is that it offers an approach designed to reduce uncertainty about the intervention's contribution to the observed results through better understanding why the observed results have occurred (or not), and the roles played by the intervention and other internal and external factors. The report from a contribution analysis is not definitive proof, but rather provides evidence and a line of reasoning from which a plausible conclusion can be drawn. This conclusion will indicate whether, within a given level of confidence, the programme has made an important contribution to the documented results.

Using programme theory to infer causation

Theory-based approaches (which include the use of theories of change, case studies, causal modelling, most significant change[7] and 'realist' reviews[8]) involve examining a particular case or set of cases in depth and theorising about the underlying causal links between actions, outcomes and impacts, thus building a programme theory of change.[9] These theories may reflect, build upon or dispute the formally accepted results framework. In relation to theories of change, the quality of evidence is related to the extent to which a credible and verifiable case can be made that the intervention contributed to the intended processes, outputs or impacts. Some newer methods for designing results frameworks, such as outcome mapping (Earl et al., 2001) and problem-driven iterative adaptation (Andrews et al., 2012), represent efforts to go beyond he mechanistic 'theories' often associated with logical frameworks. These newer methods are designed to be better aligned with monitoring and evaluation efforts that encourage critical reflection on the assumed causation in interventions.

Example of analysing contribution: NRC evaluation of advocacy and protection in the DRC

An evaluation commissioned by the Norwegian Refugee Council (NRC) assessed the outcome results of its 2012-2013 advocacy and protection initiative in the Democratic Republic of the Congo (DRC) (O'Neil and Goldschmid, 2014). It is worth noting for the following features:

- The evaluation made an attempt to estimate the level of contribution of the initiative to any changes seen at the outcome level and capture unanticipated results.
- The evaluation looked at different changes at all levels: at policy level; field level; in the area of conflict resolution; at the level of the international community; of NRC-specific field presence in DRC; and at the level of NRC programmes in general, and of the NRC advocacy programme in particular.
- The analysis of NRC's contribution to change applied indicators intended to capture and map the different levels of contribution. The evaluation team focused on providing a detailed nuance of the extent to which NRC's role and activities were visible (or not) in contributing to different changes observed.

The evaluation also used the theory of change (ToC) that was developed to inform NRC's programme in the area of access advocacy. It 'checked' the DRC's evaluation results against the ToC to gauge which piece of evaluative evidence could be situated against it to support (or not) the expected linkages between different levels of results (O'Neill and Goldschmid, 2014: 11-12).

One of the challenges was the ambition of tracking 'too many' changes (30 in total) at 'too many' levels. In the end, the evaluation team recommended reducing the number of expected changes to 10 in order to facilitate the tracking of contribution and allow for deeper analysis of fewer significant changes.

Example of mixed-method theory-based impact evaluation from WFP and UNHCR (2012)

The World Food Programme (WFP) and United Nations High Commissioner for Refugees (UNHCR) jointly commissioned and conducted a series of mixed-method impact evaluations to assess the contribution of food assistance to durable solutions in protracted refugee situations (WFP and UNHCR, 2012).

The impact evaluation series used a theory-based approach as the basis to infer causation. A synthesis evaluation was produced using a series of four standalone evaluations carried out in Bangladesh, Chad, Ethiopia and Rwanda. It used the same theoretical framework and approach, but with details adapted to the context in each case.[10] Some of the key features of this impact evaluation series are its use of:

- A logic model (visualised as a ToC) that was first developed by WFP Evaluation Office, to be then discussed and validated by the evaluation teams in the four countries visited; the teams also assessed the match of the ToC with country-level logical frameworks at different stages of the evaluation.
- A mixed-method approach to gather and analyse data in the four country cases included triangulation of data generated by desk reviews; interviews with WFP and UNHCR stakeholders; reviews of secondary data; quantitative surveys; transect walks; and qualitative interviews, including with focus groups of beneficiaries and members of local refugee-hosting communities (WFP and UNHCR, 2012: 2).
- An evaluative analysis drawing from the results that emerged from the country case studies to establish: a) which internal and external factors can causally explain the results; and b) which factors influenced the results and changes (intended and unintended) observed in the different countries and why (for more, see WFP and UNHCR, 2012: 10-13).

The evaluation team explained that the series of country-based impact evaluations were used as a way of:

> *'Test[ing] the validity of an intervention logic derived from the MOU [Memorandum of Understanding] between UNHCR and WFP and the two agencies' respective policies and programme guidance. This logic posited that the agencies' combined activities and inputs contributed to increased refugee self-reliance over three stages of evolution, starting from the refugees' situation on arrival. (…) All four [country-level impact] evaluations tested its assumptions and the extent to which food assistance contributed to outcome levels over time...' (WFP and UNHCR, 2012: 2)*

Randomised control trials

Currently, the most talked about method for assessing attribution is RCTs. In RCTs, participants are randomly assigned to a treatment (the group that received the intervention) or control group (the group that did not receive the intervention or received an alternative intervention). Provided that sampling is done carefully and that sample sizes are large enough, randomisation helps ensure that there are no systematic differences between the group that received the intervention and those who did not. This means that any observed differences may be causally attributed to the intervention. These approaches can be time-consuming and resource-intense because of the large sample size required to ensure that confounding factors are evenly distributed between treatment and control groups.[11] Proudlock, Ramalingam and Sandison (2009) cite a statement by the European Evaluation Society (2008), which proposed that RCTs should only be considered in cases where:

- a linear causal relationship can be established between the intervention and desired outcome
- it is possible to 'control' for context and other factors
- it can be anticipated that programmes under both experimental and control conditions will remain static for a considerable time
- randomness and chance have no influence
- it is ethically appropriate to engage in randomisation.

Finally, it should be stressed that RCTs should not be seen as the only 'genuine' method for assessing the attribution of impacts to an intervention. RCTs are but one tool among many for impact evaluation (Stern et al., 2012). Methods that look at contributions within a broader 'causal package' are not a 'second best' alternative to RCTs. They have different purposes and respond to different evaluation questions than RCTs.

Example of using an RCT to assess the impact of cash-based assistance: IRC in Lebanon (2014)

The International Rescue Committee (IRC) completed an evaluation of the impacts of the winter cash transfer programme run by UNHCR and partners from November 2013 to April 2014 (Lehmann and Masterson, 2014). The programme gave $575 via ATM cards to 87,700 registered Syrian refugees in Lebanon with the objective of keeping people warm and dry during the cold winter months. This was the first piece of research on causal-inference to understand the impact of conditional cash transfers and of unconditional cash transfer programmes focusing more on long-term poverty alleviation, rather than solely on immediate humanitarian needs.

This study built its sampling approach on the targeting strategy used as part of the cash assistance programme itself: cash transfers were made to benefit those living at high altitudes to target those living in the coldest areas during the winter months. In practice, those who met the threshold of living above 500 meters received cash, and those beneath 500 meters did not. The study thus compared outcomes of aid recipients residing slightly above 500 meters (treatment group who received cash) to non-recipients residing slightly below (control group who did not receive cash). It applied the same demographic criteria to calculate vulnerability, thus limiting (as much as possible) the variation between treatment group and control group.

8. Clarity around contexts and methods as the foundation for evidential quality

The degree to which it is clear why, how, and for whom evidence has been collected.

Knox-Clarke and Darcy write that:

> "Evidence, as we have seen, is information that relates to a specific proposition. As such, information is only evidence in the context of a specific question, asked by a particular organisation or group. An observer can only gauge the evidential quality of any information if they know the proposition to which the evidence relates, who wanted to prove the proposition, and how they collected the evidence. Without information about the context, it is impossible to know whether the evidence is relevant or generalisable. Similarly, information can only be accepted as evidence where the methods used to collect and analyse it, and any limitations in the exercise, are made explicit. It is only on the basis of this information that the user can determine the accuracy, representativeness, relevance, generalisability and attribution of the 'evidence'." (ALNAP, 2015)

These points can in many respects be seen to come together in the importance of being transparent and explicit about the following five questions:

- What accuracies and inaccuracies (biases, etc.) exist in the data?
- Who is 'represented' in the evidence and who is left out?
- For what the contextually derived questions can the evidence provide relevant answers?
- Which findings in the evaluation can be confidently generalised to which other contexts?
- What causal assumptions have driven the evaluation and what results or processes can be said to derive from the (often modest) contributions of the intervention?

There is no single methodological fix for these five questions. Davidson has stated that the 'gold standard' for evaluations should be about using different methods to systematically pursue causal reasoning (Davidson, 2005 and 2009). This means that realistic, reflective and contextually relevant evaluation questions should drive decisions about how to select methods that can yield quality evidence. Assuming a single superior methodological approach is to start at the wrong end.

In line with those views, Patton (2014a and 2014b) has recently shared some reflections on the questions around the quality to aim for when selecting and applying different approaches and designs to answer causal questions in evaluation. He argued that 'we need to … aim to supplant the gold standard with a new platinum standard: methodological pluralism and appropriateness.' (2014b).

Davidson (2000, 2005) suggests answering key questions in order to decide which design, method or analytical tool offers the best fit to generate quality evidence in an evaluation. This includes frank consideration of the level of certainty the evaluation commissioning agency and the intended users of the evaluation need to have, to be confident that a given intervention led to a certain change or result. Many decision-makers are prepared to make decisions if they are, say, 70% or 80% certain of the evidence provided to prove or disprove a given evaluative statement. Different contexts and different types of decision will call for different 'thresholds' of certainty. And indeed, because each decision context requires a different level of certainty, it is important to

be clear upfront about the level of certainty required by decision-makers and other evaluation stakeholders (Davidson, 2005: 69). The depth and breadth of the required evidence base is a key consideration in evaluation planning and should be based on a transparent discussion of stakeholder information needs. This will help not only with appropriately budgeting the evaluation, but will facilitate upfront discussions about the trade-offs between budgets, timelines, and the confidence level of conclusions (Davidson, 2000: 25).

Having some clarity on these types of stakeholders' expectations should thus help evaluators better explain the implications for the quality of evidence inherent in their design and method choices. It should also help evaluators engage in a concrete discussion of the trade-offs in decisions about the levels and types of quality the evaluation requires. The ultimate goal – or platinum standard, to use Patton's words – is to identify and propose the most appropriate blend or best fit of designs and methods to answer the evaluation questions at hand.

9. Questions for design and dialogue around the quality of evidence

The following questions are a proposed basis for discussions around how to promote better quality evidence in humanitarian evaluations. They are not intended as a set 'checklist' for how to 'do things right', but are intended to provide guidance in considering options and increasing awareness of important issues. These questions are expected to be useful for evaluation managers, those commissioning evaluations and evaluation teams in discussing the following issues:

- Decisions about the scope and resources allocated for evaluations (in connection with the intended uses of the evaluation and expected confidence in evidence).
- Choices in design of the evaluation methodologies and later discussions of how to analyse findings.
- Guidance in assessing the quality of evaluations, though it is acknowledged that the extent to which evaluation reports describe how they have addressed these issues, is likely to vary.

Furthermore, the questions are presented to encourage thinking on how to promote sharing and inclusion of people affected by crises and partners in local research institutions in evaluation processes. By being transparent about our criteria for what constitutes 'good evidence', we can establish a basis for constructive dialogue about how to bring in evidence from different sources and perspectives. Ideally, this can broaden the discussion of evaluation quality to better embrace the range of perceptions of evidence quality within the humanitarian evaluation community and among the users of evaluations.

The questions below are structured around the six ALNAP criteria on quality of evidence:

Accuracy

- How can the evaluation design present a 'true' picture that triangulates both 'objective facts' regarding the intervention and also subjective perceptions, while recognising that each are more or less appropriate depending on the evaluative question being posed?
- Has the evaluation design and implementation considered risks of bias by both informants and evaluators, and proposed possible measures to minimise (or at least acknowledge) these risks in methods and subsequent analyses?
- Are the motives of respondents in presenting a certain impression of the intervention acknowledged?
- Have steps been taken to triangulate or to find ways to look critically at and reassess data when informants show signs of 'telling us what we want to/expect to hear'?
- When relying on triangulation of evidence, is the evaluation transparent about the weighting and confidence levels of different types and sources of evidence?
- Where triangulation leads to findings that highlight the contradictory or contrasting perspectives on an issue, is this acknowledged and analysed?
- Where baseline or monitoring data is absent, skewed or focus on 'the wrong indicators', is this addressed (through exploration of other data sources)? At a minimum, is it used to inform stakeholders so as to encourage improvements in future monitoring efforts?

Representativeness

- When a purposive sample of a set of stakeholders is being used, is the sample based on an analysis of the different 'stakes' that the respondents have in the intervention, and whose voices they 'represent'? Is this anchored in the ToC of the intervention in relation to outcomes and impacts on different parts of the population?
- Are choices about how to optimise sampling methods in relation to timeframes, access and available resources made transparent? Are the implications of these choices transparently considered in relation to the purposes and ambition levels of the evaluation?
- How has the evaluation judged 'saturation' i.e. ensuring that they have captured a full picture of the situation pertaining to the evaluation questions?
- Does the evaluation require an accurate overview of the effects of the intervention in relation to the overall population? Or is it more concerned with specific categories of intended aid recipients/programme participants, service users or other stakeholders? How has that been reflected in the choice of sampling methods?
- Is the evaluation clear about whose voices it intends to 'represent' in the data collected (and whose voices are excluded)?

Relevance

- Is the evaluation explicit about the indicators against which the intervention is being assessed, while also taking a critical stance as to whether those indicators really reflect the propositions or policies being interrogated?
- Has the evaluation been designed so as to critically assess the underlying propositions (explicit and implicit) in the ToC of the intervention?
- Has the evaluation design considered participatory methods to draw attention to the indicators that the people affected by the conflict or disaster use to judge the quality of the intervention?
- Have steps been taken to encourage ownership of evaluation findings and recommendations among different sets of decision-makers (including field level staff and service providers) by ensuring questions and indicators used are relevant for providing them with advice and guidance for the decisions they need to make?
- When using participatory methods, has the evaluation team been given sufficient independence from those directly engaged in the intervention to ensure that the situation and needs of 'non-participating' population groups are analysed?

Generalisability

- Are the findings of the evaluation meant to be generalisable to other contexts?
- Has the evaluation effort exercised due caution in judging what can be generalised from the findings to other contexts i.e. by making explicit what is unique to a given country, conflict or convergence of risks?
- Has the evaluation effort exercised due caution in judging what can be generalised from the findings to other interventions, i.e., by making explicit what is unique to a given type of programme, modality or method?

- Does the evaluation have a structured and transparent method for determining what can and cannot be generalised from one case to another? Does it describe the 'patterns' in the underlying dimensions of the interventions that are similar in other interventions and contexts?
- Has the evaluation considered how much confidence in the evidence is needed to conclude that findings from a given evaluation can be judged to be relevant for informing the use of a given approach in a different context?

Attribution

- Are assumptions about attribution made transparent and clear? Are they supported by acceptable methods?
- Are there any doubts about what can actually attributed to the intervention explained, and alternative explanations of contribution transparently interrogated, where appropriate?
- Does the evaluation acknowledge (and even highlight) the broader contextual factors and other related interventions that together impinge on the likelihood that the intervention contributed to the results claimed?
- Has the evaluation method and process provided space for recognising unintended positive or negative effects of the intervention?
- Has the selection of methods used control or quasi-control design to establish a causal relationship? Has it reflected the range of causal factors and intended results that the evaluation has been tasked with assessing?
- Have the methods been applied in a sufficiently rigorous manner to either indicate the strength of the causal relationship or disentangle the causal package in a verifiable manner?

Clarity around context and methods

- Are the basic facts surrounding the evaluation presented i.e. who commissioned the evaluation, when and why it was commissioned, who conducted the evaluation and what were the methods and limitations?
- Does the evaluation design reflect an optimal approach for answering the evaluation questions in a rigorous and verifiable manner given the resources available? Have the methodological choices for the evaluation been explained, including the advantages and constraints of alternative approaches?
- Is the evaluation report transparent about: judgements regarding the accuracy of the data; whose perspectives are included (and excluded) in the findings; and what the underlying assumptions are behind the theory of change of the intervention?
- Have stakeholders (ideally those being targeted or otherwise affected by the intervention) had an opportunity to validate the findings and conclusions of the evaluation?
- Have the selection of methods and decisions about investment of resources for the evaluation reflected the threshold of certainty needed to use and learn from the evaluation?
- Has the design of the evaluation (including inception discussions) served to focus the evaluation on questions that are realistic to answer, critically reflective and contextually relevant?

Endnotes

1. Overview of the meeting, including all presentations and panel material and videos are available on the ALNAP website: **www.alnap.org/events/28th**
2. Method notes are available at: **http://www.alnap.org/resources/results.aspx?q=eha%20method%20note%7Cta**
3. **Primary data** refers to data gathered directly by the Evaluation Team e.g. in their interactions with crisis affected populations, other people of concern, and other various stakeholders in partner agencies and government entities. **Secondary data** is gathered from documents, reports, analysis produced by actors/entities other than the Evaluation Team.
4. **See: http://betterevaluation.org/evaluation-options/triangulation**
5. The reasons for this 'path dependency' are many, but quite often relate to fears of donor displeasure if initial plans are not executed as intended.
6. For an introduction on this specific issue, see Gerring, 2012.
7. The BetterEvaluation portal offers specific guidance materials on all these approaches: **http://betterevaluation.org/plan/understandcauses/check_results_match_theory**
8. For a short and accessible introduction, see Westhorp (2014). For an example of how realist reviews are conducted, see Westhorp et al. (2014).
9. For an overview, see Rogers, 2000.
10. The series of reports is available from the WFP Evaluation Office: **http://www.wfp.org/evaluation/**
11. For more on this, see Jones, 2009.

Bibliography

Abebe, D. and Catley, A. (2012) 'Participatory impact assessment in drought policy contexts: lessons from southern Ethiopia', in Holland, J. (ed.) *Who counts? The Power of Participatory Statistics.* Rugby: Practical Action Publishing, pp. 149-162. **http://www.alnap.org/resource/13030**

ACAPS (2013) *How sure are you? Judging quality and usability of data collected during rapid needs assessments*. (Technical Brief). Geneva: Assessment Capacities Project. **http://www.alnap.org/resource/11437**

Andrews, M. Pritchett, L. and Woolcock, M. (2012) *Escaping capability traps through Problem-Driven Iterative Adaptation* (PDIA). Center for Global Development Working Paper 299, June. **http://www.alnap.org/resource/24292**

Ager, A., and Metzler, J. (2012) *Child Friendly Spaces. A Structured Review of the Current Evidence Base.* Columbia University Mailman School of Public Health and World Vision International. New York and Geneva. **http://www.alnap.org/resource/19063**

Alexander, J. (2014) 'Improving the quality of EHA evidence'. *ALNAP Discussion Series, EHA Method note no 2.* London: ALNAP/Overseas Development Institute. **http://www.alnap.org/resource/12636**

Alexander, J. and Cosgrave, J. (2014) 'Representative sampling in humanitarian evaluation'. *ALNAP Discussion Series, EHA Method note no 1*. London: ALNAP/Overseas Development Institute. Available: **http://www.alnap.org/resource/10389**

Annan, J. (2014) 'Mapping the evidence base in conflict and post-conflict contexts'. *IRC blog*. Posted on April 18. Available at: **http://www.rescue.org/blog/mapping-evidence-base-conflict-and-post-conflict-contexts**

Assessment Working Group for Northern Syria (2012) *Joint Rapid Assessment of Northern Syria: Aleppo City Assessment - March 2013*. **http://www.alnap.org/resource/8755**

Aubel, J. (1999) *Participatory program evaluation manual*. Calverton and Baltimore: Child Survival Technical Support Project and Catholic Relief Service. **http://www.alnap.org/resource/13023**

Bamberger, M., Rugh, J., Mabry, L. (2012) *Real World Evaluation – Working under budget, time, data and political constraints.* (2nd edition) SAGE Publications. **http://www.alnap.org/resource/8076**

Beach, D. and Pedersen, R. (2011) *What is process tracing actually tracing? The three variants of process tracing methods and their uses and limitations*. Aarhus: Aarhus University. **http://www.alnap.org/resource/22126**

Befani, B. (2012) 'Models of causality and causal inference', in Stern et al. DFID Working Paper 38, *Broadening the Range of Designs and Methods for Impact Evaluations*. London: Department for International Development. **http://www.alnap.org/resource/8196**

Blatter, J. and Blume, T. (2008) 'In search of co-variance, causal mechanisms or congruence? Towards a plural understanding of case studies'. *Swiss Political Science Review,* 14(2): 315–356. **http://www.alnap.org/resource/22131**

Brikci, N. and Green, J. (2007) *A guide to using qualitative research methodology*. London: MSF and London School of Hygiene and Tropical Medicine. **http://www.alnap.org/resource/13024**

Brown D., Donini, A. and Knox-Clarke, P. (2014) 'Engagement of crisis-affected people in humanitarian action'. *Background Paper of ALNAP's 29th Annual Meeting*, 11-12 March, Addis Ababa. London: ALNAP/ Overseas Development Institute. **http://www.alnap.org/resource/10439**

Brown, D. and Donini, A. (2014) *Rhetoric or reality? Putting affected people at the centre of humanitarian action*. ALNAP Study. London: ALNAP/Overseas Development Institute. **http://www.alnap.org/resource/12859**

Buchanan-Smith, M. and Cosgrave, J. (2013) *Evaluation of humanitarian action: Pilot guide*. London: ALNAP/ Overseas Development Institute. **http://www.alnap.org/resource/8229**

Cartwright, N. (2010) *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge, UK: University Press. **http://www.alnap.org/resource/24293**

Cartwright, N. and Hardie, J. (2012) *Evidence-based policy: doing it better. A practical guide to predicting if a policy will work for you*. Oxford, UK: Oxford University Press. **http://www.alnap.org/resource/22134**

Catley, A., Burns, J., Abebe, D. and Suji, O. (2013) *Participatory impact assessment: A design guide.* Somerville: Feinstein International Center, Tufts University. **http://www.alnap.org/resource/10811**

Chambers, R. (2002) *Participatory workshops: A sourcebook of 21 sets of ideas and activities*. London: Earthscan Publications Ltd. **http://www.alnap.org/resource/11838**

Chambers, R. et al. (2009) *Designing impact evaluations: different perspectives*. 3ie Working Paper no.4. New Delhi: International Initiative for Impact Evaluation. **http://www.alnap.org/resource/8474**

Chan, J. (2014) 'Oxfam America's community-based Drought Early Warning Surveillance approach'. Presented at *Engagement of Crisis-affected People in Humanitarian Action*, ALNAP 29th Annual Meeting, Addis Ababa, 11-13 March. **http://www.alnap.org/resource/10784**

Chigas, D., Church, M., Corlazzoli, V. (2014) 'Evaluating Impacts of Peacebuilding Interventions'. *Practice Products for the Conflict, Crime, and Violence Results Initiative* (CCVRI) Dfid. London: Dfid. **http://www.alnap.org/resource/19072**

Coe, J. and Majot, J. (2013) *'Monitoring, Evaluation and Learning (MEL) in NGO advocacy - Findings from Comparative Policy Advocacy MEL Review Project'*. Boston, MA: Oxfam America. **http://www.alnap.org/resource/19073**

Cook, T. (2000) 'The false choice between theory-based evaluation and experimentation'. *New Directions for Evaluation*, Vol. 87 pp. 27-34. **http://www.alnap.org/resource/22322**

Cook, T. et al. (2010) 'Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven'. *American Journal of Evaluation*, 31(1), 105-117. **http://www.alnap.org/resource/22323**

Corlazzoli, V. and White, J. (2013) 'Back to Basics - A compilation of best practices in design, monitoring & evaluation in fragile and conflict-affected environments'. *Practice Products for the CCVRI – Conflict, Crime, and Violence Results Initiative, Dfid*. London: Dfid. **http://www.alnap.org/resource/10196**

Cornwall, A. (2014) *'Using participatory process evaluation to understand the dynamics of change in a nutrition education programme*'. IDS Working Paper no 437. Brighton: Institute of Development Studies. **http://www.alnap.org/resource/22891**

Coryn, C. (2013) *Experimental and quasi-experimental designs for evaluation*. Washington DC: American Evaluation Association. Training materials. **http://www.alnap.org/resource/22326**

Cosgrave, J. (2015) Quality control in evaluation. Posted on the ALNAP EHA Community of Practice on 16 June. Available at: **https://partnerplatform.org/alnap/humanitarian-evaluation/discussions/8d1bb007**

Court, J., & Young, J. (2003) '*Bridging Research and Policy: Insights from 50 Case Studies Working Paper'.* London: Overseas Development Institute. **http://www.alnap.org/resource/8100**

Cousins, J. B. and Whitmore, E. (1998) 'Framing participatory evaluation'. *New Directions for Evaluation*, 1998(80): 5-23. **http://www.alnap.org/resource/13028**

Cullen, A. and Coryn, C. (2011) 'Forms and functions of participatory evaluation in international development: A review of the empirical and theoretical literature'. *Journal of MultiDisciplinary Evaluation*, 7(16): 32-47. **http://www.alnap.org/resource/13026**

Daniel, J. (2012) *Sampling Essentials: Practical guidelines for making sampling choices.* Washington, DC: Sage Publications. **http://www.alnap.org/resource/9161**

Darcy, J, Stobaugh, H., Walker, P. and D. Maxwell (2013) '*The Use of Evidence in Humanitarian Decision Making*'. ACAPS Operational Learning Paper. Boston: Feinstein International Center and Tufts University. **http://www.alnap.org/resource/8003**

Davidson, J. (2000) 'Ascertaining causality in theory-based evaluation'. *New Directions for Evaluation, Special Issue: Program Theory in Evaluation: Challenges and Opportunities.* Vol. 2000 (87) pp. 17–26. **http://www.alnap.org/resource/19075**

Davidson, J. (2005) *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation*. Thousand Oaks: Sage. **http://www.alnap.org/resource/19076**

Davidson, J. (2009) 'Causal inference: Nuts and bolts'. Presentation for ANZEA evaluation conference. **http://www.alnap.org/resource/22908**

Davidson, J. (2013) 'Understand Causes of Outcomes and Impacts'. American Evaluation Association (AEA) Coffee Break Demonstration CBD141 delivered on 21 March. **http://www.alnap.org/resource/19077**

Department for International Development (2012) *Broadening the range of designs and methods for impact evaluations*. London: DFID. **http://www.alnap.org/resource/8196**

Earl, S, Carden, F. and Smutylo, T. (2001) *Outcome Mapping: Building learning and reflection into development programs*. International Development Research Centre, Ottawa. **http://www.alnap.org/resource/11835**

Epstein, L. (2013) 'The Whole (Behavioral) Truth and Nothing But the Truth? Not Likely…' *Survey Monkey Blog* post, 24 April. **http://www.alnap.org/resource/24294**

EES (European Evaluation Society) (2008) *ESS Statement: The Importance of a Methodologically Diverse Approach to Impact Evaluation – Specifically with Respect to Development Aid and Development Interventions.* December, EES. **http://www.alnap.org/resource/11119**

Foley, P. (2009) *Participatory evaluation of the 2008 farmer field school programme, Lira, Uganda.* London: ACF. **http://www.alnap.org/resource/20908**

George, A. L. and Bennet, A. (2004) 'The Method of structured, Focused Comparison', in *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press. **http://www.alnap.org/resource/22152**

Gerring, J. (2011) *Social Science Methodology A Unified Framework*. (2nd edition) Cambridge: Cambridge University Press. **http://www.alnap.org/resource/9015**

Goertz, G. and Mahoney, J. (2012) *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton University Press. **http://www.alnap.org/resource/13029**

Goldwyn, S. and Chigas, D. (2013) *Monitoring and evaluating conflict sensitivity – methodological challenges and practical solutions.* Practice Product developed under the Conflict, Crime, and Violence Results Initiative (CCVRI). London: Dfid. **http://www.alnap.org/resource/22338**

Ground Truth Solutions (n.d.) *Social Science meets customer satisfaction*. Ground Truth portal information post. Available: **http://groundtruthsolutions.org**

Guest, G., Bunce, A., & Johnson, L. (2006) *How many interviews are enough?* Field methods, 18(1), 59-82. **http://www.alnap.org/resource/22952**

Guijt, I. and Gaventa, J. (1998) 'Participatory monitoring & evaluation: Learning from change'. *IDS Policy Briefing Issue no 12.* Brighton: Institute of Development Studies. **http://www.alnap.org/resource/13027**

Hallam, A. and Bonino, F. (2013) *Using evaluation for a change*. ALNAP Study. London: Overseas Development Institute/ALNAP. **http://www.alnap.org/resource/8980**

Holland, J. (2013) *Who counts? The power of participatory statistics*. Rugby, UK: Practical Action Publishing. **http://www.alnap.org/resource/13030**

Holzmann, P. with Boudreau, T., Holt, J., Lawrence, M. and O'Donnell, M. (2008) *The household economy approach: A guide for programme planners and policy-makers*. London: Save the Children. **http://www.alnap.org/resource/13036**

Hughes, K. and Hutchings, C. (2011) *Can we obtain the required rigour without randomisation? Oxfam GB's non-experimental Global Performance Framework.* 3ie Working Paper no. 13. New Delhi: International Initiative for Impact Evaluation (3ie) **http://www.alnap.org/resource/24295**

IFRC-PED (2014) 'PMER training materials'. Geneva: IFRC Planning and Evaluation Department (PED).

Jones, H (2009) 'The "gold standard" is not a silver bullet for evaluation'. Opinion Piece. London: Overseas Development Institute. **http://www.alnap.org/resource/11123**

Karlan, D. (2009) 'Thoughts on Randomized Trials for Evaluation of Development: Presentation to the Cairo Evaluation Clinic', in Chambers, et al. (2009) *Designing impact evaluations: different perspectives.* 3ie Working Paper no.4. New Delhi: International Initiative for Impact Evaluation. pp. 8-13.
http://www.alnap.org/resource/8474

Kielmann, K., Cataldo, F. and Seeley, J. (2011) *Introduction to qualitative research methodology.* London: DfID.
http://www.alnap.org/resource/6264

King, J. (2005) 'Participatory evaluation', in Mathison, S. (ed.) *Encyclopedia of evaluation*. London and New York: Sage Publications. **http://www.alnap.org/resource/22667**

King, J. A., Cousins, J. B. and Whitmore, E. (2007) 'Making sense of participatory evaluation: Framing participatory evaluation'. *New Directions for Evaluation,* 2007(114): 83-105
http://www.alnap.org/resource/13032

Knox-Clarke, P. and Darcy, J. (2014) *Insufficient evidence? The quality and use of evidence in humanitarian action.* ALNAP Study. London: ALNAP/ Overseas Development Institute.
http://www.alnap.org/resource/10441

Lansdown, G. and O'Kane, C. (2014) *A toolkit for monitoring and evaluating children's participation: A 10-step guide to monitoring and evaluating children's participation*. London: Save the Children.
http://www.alnap.org/resource/11639

Lehmann, C. and Masterson, D. (2014) *Emergency Economies: The Impact of Cash Assistance in Lebanon. An Impact Evaluation of the 2013-2014 Winter Cash Assistance Program for Syrian Refugees in Lebanon*. Beirut, Lebanon: IRC - International Rescue Committee. **http://www.alnap.org/resource/12921**

Longhurst, R. (2013) *Implementing development evaluations under severe resource constraints.* CDI Practice Paper. Brighton: IDS. **http://www.alnap.org/resource/8950**

Mathison, S. (2014) 'Participatory evaluation', in Coghlan, D. and Brydon-Miller, M. (eds.) *The SAGE encyclopedia of action research*. Thousand Oaks: Sage. **http://www.alnap.org/resource/13033**

Mathison, S. (ed.) (2005) *Encyclopedia of Evaluation*. London and New York: Sage.
http://www.alnap.org/resource/22667

Mayne, J. (2012) *Making causal claims*. ILAC Brief No. 26. Rome: Institutional Learning and Change (ILAC) Initiative. **http://www.alnap.org/resource/24296**

Meier, P. (2012) 'Humanitarian action 2.0', in the *Magazine of the Red Cross and Red Crescent Movement*. **http://www.alnap.org/resource/24297**

Mendizabal, E. (2010) *The Alignment, Interest and Influence Matrix* (AIIM). London: Overseas Development Institute/RAPID. **http://www.alnap.org/resource/19252**

NAO (2000) *A practical guide to sampling*. London: National Audit Office.
http://www.alnap.org/resource/10228

Nutley, S., Powell, A. and Davies, H. (2012) *What counts as good evidence? A provocation prepared for the Alliance for Useful Evidence.* London: NESTA. **http://www.alnap.org/resource/9606**

O'Neil, G. (2014) 'Discussing accuracy in humanitarian evaluations – Round-up for final comments' [ALNAP Humanitarian Evaluation Community of Practice]. 14 May.
Available: **https://partnerplatform.org/alnap/humanitarian-evaluation/discussions/61469506**

O'Neil, G. and Goldschmid, P. (2014) *Final Report Evaluation of NRC's 2012-13 protection and advocacy work in the DRC*. Oslo: Norwegian Refugee Council. **http://www.alnap.org/resource/24298**

OCHA (2014) 'Participative evaluation of accountability to affected populations: Central African Republic, version 2.0'. Bangui: OCHA Country Office. **http://www.alnap.org/resource/13034**

Patton, M. Q. (2011) *Developmental evaluation: Applying complexity concepts to enhance innovation and use.* New York: The Guildford Press. **http://www.alnap.org/resource/8499**

Patton, M. Q. (2012) 'Contextual Pragmatics of Valuing'. *New Directions for Evaluations,* Vol 2012 (133), 1-129. **http://www.alnap.org/resource/24299**

Potts, A., Mayers, K., and Roberts, L. (2011) 'Measuring human rights violations in a conflict-affected country: results from a nationwide cluster survey in Central African Republic'. *Conflict and Health* Vol (5) 4, 14. **http://www.alnap.org/resource/12629**

Pritchett, L. and J. Sandefur (2013) *Context Matters for Size: Why external validity claims and development practice don't mix*. Center for Global Development. **http://www.alnap.org/resource/9164**

Puri, J. (2013) 'Experimental methods for impact evaluation'. Presentation at ALNAP's 'Skills-building day for evaluators' on 14 March, Washington, DC. **http://www.alnap.org/resource/8038**

Puri, J., Aladysheva, A., Iversen, V., Ghorpade, Y. and Brück, T. (2014) *What methods may be used in impact evaluations of humanitarian assistance?* 3ie Working Paper 22. New Delhi: International Initiative for Impact Evaluation. **http://www.alnap.org/resource/19288**

Read, M. (2014) 'Are we sure? Veracity in Humanitarian Evaluations' [ALNAP Humanitarian Evaluation Community of Practice]. 7 April.
Available: **https://partnerplatform.org/alnap/humanitarian-evaluation/discussions/7c1ca59b**.

Robert Wood Johnson Foundation (2008) *Qualitative research guidelines project*. **http://www.qualres.org/**

Rogers, P, J. (2000), Causal models in program theory evaluation. *New Directions for Evaluation*, 87, 47-55. **http://www.alnap.org/resource/19116**

Ruzzene, A. (2014) *Using case studies in the social sciences. Methods, inferences, purposes*. Unpublished Doctoral thesis. University of Rotterdam, Netherlands.

Save the Children (2000) 'Participatory monitoring and evaluation: Methodologies for working with children and young people'. Save the Children briefing. London: Save the Children.

Seaman, J., Clarke, P., Boudreau, T. and Holt, J. (2000) *The Household Economy Approach: A Resource Manual for Practitioners*. London: Save the Children. **http://www.alnap.org/resource/10029**

Segone, M. and Bamberger, M. (2011) *How to design and manage equity-focused evaluations*. New York: UNICEF. **http://www.alnap.org/resource/8080**

Shannon, H. et al. (2012) '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*, Vol. 9:5. **http://www.alnap.org/resource/8156.aspx**

Stocké, V. and Langfeldt, B. (2004) 'Effects of Survey Experience on Respondents' Attitudes Towards Surveys'. *Bulletin of Sociological Methodology*, Vol. 81, 5-32. **http://www.alnap.org/resource/12628**

Tsui, J., Hearn, S. and Young, Y. (2014) *Monitoring and evaluation of policy influence and advocacy*. ODI Working Paper no. 395. London: Overseas Development Institute. **http://www.alnap.org/resource/12943**

UNICEF (2002) 'Children participating in research, monitoring and evaluation (M&E): Ethics and your responsibilities as a manager'. Evaluation technical note no 1. New York: UNICEF Evaluation Office. **http://www.alnap.org/resource/8174**

Weaver, L. and Cousins, J. B. (2004) 'Unpacking the participatory process'. *Journal of MultiDisciplinary Evaluation*, 1(1): 19-40. **http://www.alnap.org/resource/13037**

Westhorp, G. (2014) 'Realist impact evaluation: an introduction'. *Methods Lab*. London: Overseas Development Institute. **http://www.alnap.org/resource/19141**

Westley, K. and Mikhalev, V. (2002) *The use of participatory methods for livelihood assessment in situations of political instability: A case study from Kosovo*. Working Paper no 190. London: Overseas Development Institute. **http://www.alnap.org/resource/6899**

WFP and UNHCR (2013) *Synthesis Report of the Joint WFP and UNHCR Impact Evaluations on the Contribution of Food Assistance to Durable Solutions in Protracted Refugee Situations.* Rome and Geneva: WFP-Evaluation Office and UNHCR/PDES. **http://www.alnap.org/resource/12260**

White, H. and D. Philips (2012) *Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework*. 3IE. **http://www.alnap.org/resource/8205**

Yin, R. (2010) 'Analytic Generalization', in Mills, A. J., Durepos, G. and Wiebe, E. (eds.) *Encyclopedia of Case Study Research*. Thousand Oak: Sage. **http://www.alnap.org/resource/24300**

Yin, R. (2008) *Case Study Research: Design and Methods*. London: Sage. **http://www.alnap.org/resource/9006**

Young, J. et al. (2014) RAPID *Outcome Mapping Approach (ROMA) - a guide to policy engagement and influence.* London: Overseas Development Institute/RAPID. **http://www.alnap.org/resource/1294**

Other ALNAP resources on evaluation

Evaluating Humanitarian Action Guide

Evaluating protection in humanitarian action:
Issues and challenges

Using evaluation for a change:
Insights from humanitarian practitioners

EHA Method Notes on representativeness,
accuracy, evidence and causation

EHA Practice Note:
Repeat after me: Communicate,
disseminate and support take-up!

Insufficient evidence?
The quality and use of evidence in humanitarian action

Real-time evaluations of humanitarian action:
An ALNAP guide

Evaluating Humanitarian Action
using the OECD-DAC Criteria:
An ALNAP Guide for humanitarian agencies

www.alnap.org/what-we-do/evaluation

ALNAP
Overseas Development Institute
203 Blackfriars Road
London SE1 8NJ
United Kingdom
Email: alnap@alnap.org