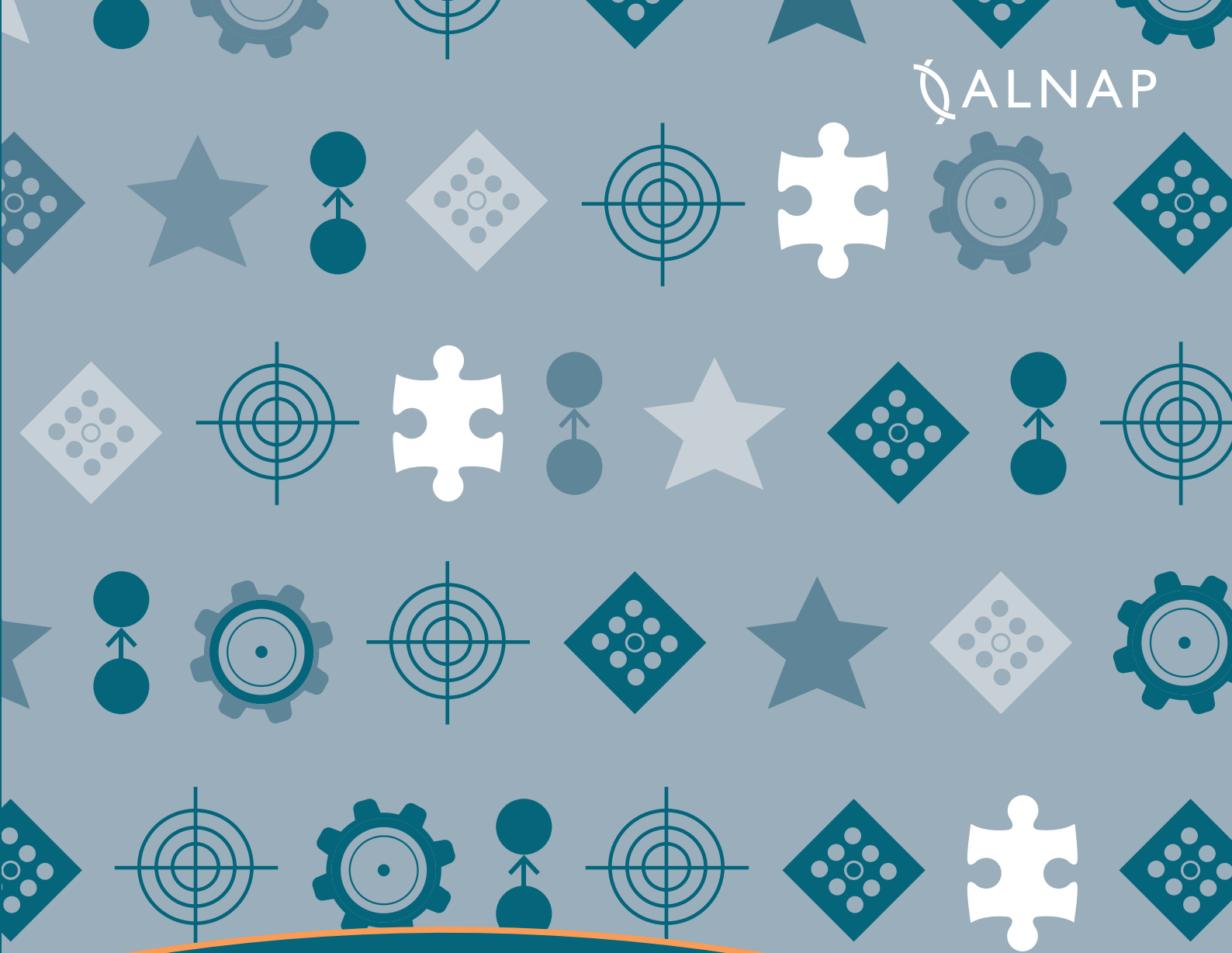


INSUFFICIENT EVIDENCE?

The quality and use of evidence in humanitarian action

Paul Knox Clarke and James Darcy

ALNAP STUDY

ALNAP is a unique system-wide network dedicated to improving the performance of humanitarian action through shared learning.
www.alnap.org

Have you read the **ALNAP Discussion Starter** that accompanies this study?
www.alnap.org/pool/files/discussion-starter-evidence-alnap-2014.pdf

An electronic copy of this study, the discussion starter and other related resources are available on the ALNAP website at www.alnap.org/ourwork/evidence.

Suggested citation
Knox Clarke, P. and Darcy, J. (2014) *Insufficient evidence? The quality and use of evidence in humanitarian action*. ALNAP Study. London: ALNAP/ODI.

© ALNAP/ODI 2014. This work is licensed under a Creative Commons Attribution-NonCommercial Licence (CC BY-NC 3.0).

ISBN 978-0-9926446-4-2

Copy-edited by Angela Hawke
Typeset by Soapbox, www.soapbox.co.uk

Contents

Acknowledgements 3

Introduction 5

1.1 Why does evidence matter for humanitarian action? 5
1.2 About this paper 6

What is 'evidence' in the context of humanitarian action and how do we judge its quality? 10

2.1 Background: terms, concepts and challenges 10
2.2 Evidence for what? Humanitarian propositions and evidence requirements 12
2.3 Criteria to assess the quality of humanitarian evidence 15

Does humanitarian evidence currently meet the criteria for quality? 19

3.1 The challenges to generating evidence in humanitarian situations 19
3.2 Evidence from early warning systems 21
3.3 Evidence from needs assessment 23
3.4 Evidence from monitoring 31
3.5 Evidence from evaluations and controlled trials 38

Is evidence currently used to guide humanitarian operations and policy-making? 46

4.1 Quality does not guarantee use 46
4.2 Use of early warning evidence 46
4.3 Use of assessment and monitoring evidence 50
4.4 Use of evaluation evidence 54

Conclusions and recommendations: how can the quality and use of evidence in humanitarian action be improved? 67

5.1 Improving the quality of evidence 67
5.2 Improving the use of evidence in humanitarian decision-making 69

Annex 1: Methodology 73

Annex 2: List of interviewees and contributors 74

Bibliography 75

Acknowledgements

This paper was produced by the ALNAP Secretariat and builds on the paper *Evidence and Knowledge in Humanitarian Action*, produced for the 28th ALNAP Annual Meeting (March 2013). The original paper was written by James Darcy and Paul Knox Clarke, with research assistance from Miriam Bradley and Alexandra Warner. Leah Campbell provided research assistance to the current study. The authors would like to express their thanks to all those who participated in the Annual Meeting, and particularly to those who made presentations, whose experiences and insights form an important element of this revised paper. The authors would also like to thank all those who agreed to be interviewed for this paper, whose names can be found in Annex 2. Finally, the authors are indebted to Richard Garfield, Joanna Macrae and Anthony Redmond, who, as peer reviewers, made suggestions that have improved the quality of this report significantly. The authors remain fully responsible for the views expressed and for any errors or omissions in the text.

Introduction

1.1 Why does evidence matter for humanitarian action?

The failure to generate and use evidence in policy and response makes humanitarian action less effective, less ethical and less accountable.

Effective humanitarian responses depend as much on knowledge as they do on funding or logistics. Indeed, without knowledge, responses cannot hope to succeed: 'The international humanitarian community's ability to collect, analyse, disseminate and act on key information is fundamental to effective response' (Mahmood et al., 2010).

“Effective humanitarian responses depend as much on knowledge as they do on funding or logistics.”

Equally importantly, in a sector rooted so firmly in values, evidence matters if we are to uphold humanitarian principles: 'If you believe in impartiality, you have to be evidence based. You can't be impartial if you don't know what the range of choices are' (Peter Walker, personal communication, March 2013). A recent DFID study on the challenges of collecting evidence in emergencies highlights the need to follow ethical frameworks when doing so and also states that it is 'unethical to deliver interventions that are at best not proven, are ineffective or, worse still, do actual harm' (DFID, 2012: 11).

Finally, evidence matters for accountability. Donor organisations, affected states and civil society have a legitimate expectation to know how – and how well – money is spent on behalf of people caught up in crises. To meet these expectations, humanitarian organisations must be able to prove that needs exist and demonstrate that they have made informed and deliberate choices about the most effective and efficient way to respond. They must also be able to provide evidence on the impact of the choices they make.

This paper (and the ALNAP meeting on which it is based) is underpinned by the sense that 'at present, humanitarian decisions are often based on poor information' (DFID, 2012: 5) and are 'anecdote, rather than evidence, driven' (Mazurana et al., 2011: 1). Even when evidence is available, decisions can still appear to be driven by personal conviction, or by political or fundraising considerations. So it is important that the evidence available be of the highest quality, but it is equally important that this evidence should be *used* by decision-makers. Recurrent collective failures to respond decisively in the face of strong evidence of impending crises (notably from famine early warning systems in sub-Saharan Africa) confirm that *generating* such evidence is only one part of the challenge. The same is true for evidence from past experience: one recurrent theme of evaluations is that the international system and its individual organisations struggle to learn the lessons of the past and apply its evidence to today's humanitarian practice (Sandison, 2006; Hallam, 2011).

Such concerns are not unique to humanitarian action. Many areas of public policy – including health, education and international development – have, in the past 15 years, recognised the need to integrate evidence more systematically into practice. Stern et al. (2012) describe the rise of the Evidence Based Policy (EBP) movement, which argues that policy-makers should take decisions based on evidence rather than on ideology or in response to special interests (see also Davies et al., 2000; Nutley et al., 2007; NAO, 2003). This stronger focus on evidence has generated debate over the very meaning of the term, over what counts as 'evidence' and over the best approaches to generating evidence. This paper considers some of these questions in the specific context of the international response to humanitarian emergencies.

The generation and use of evidence in humanitarian action is not a new topic: discussions on evidence were part of the push for greater quality and accountability in humanitarian action in the 1990s. Since then there have been a number of technical initiatives to improve the quality of information provided by early warning, assessment and evaluation mechanisms, aiming, in general, to ensure that humanitarian responses are genuinely 'needs based', and founded on evidence of 'what works'.

There have also been attempts to ensure the wider use of evidence in humanitarian action. Many organisations have tried to strengthen their own systems for organisational learning and knowledge management, while the system as a whole has supported collective endeavours for joint learning, codification and standard-setting (Walker and Purdin, 2004; Young and Harvey, 2004; Barnett, 2005). However, there is still some way to go in improving both the quality and the use of evidence in the humanitarian sector.

IN BRIEF, THE PAPER FINDS THAT:

- despite progress over the past 20 years, there appears to be room for improvement in the quality and use of evidence in international humanitarian action;
- evidence matters: the use of good quality evidence improves the effectiveness and accountability of humanitarian action, and is in accordance with humanitarian ethics and principles;
- discussions on evidence in humanitarian action can be seen in the context of broader discussions around the generation and use of evidence in public policy.

1.2 About this paper

This paper reviews the quality of evidence available today to support humanitarian action. It focuses primarily on evidence generated by the 'formal international humanitarian system'[1] through early warning, needs assessment, monitoring and feedback, evaluation and impact assessment. It also considers the degree to which actors in the humanitarian system actually use evidence to guide their operational

1 A term used to describe national and international NGOs, UN agencies, members of the Red Cross/Crescent family, and host and donor government entities involved in funding, programming and implementing international humanitarian assistance. The concept of formal humanitarian system used in this paper follows that used in ALNAP's State of the Humanitarian System: see (ALNAP, 2012: 16).

decisions and policy formulation. The paper proposes the use of six main criteria for the quality of evidence:

- Accuracy
- Representativeness
- Relevance
- Generalisability of conclusions
- Attribution
- Clarity around context and methods.

The paper is aimed at humanitarian policy-makers, humanitarian decision-makers (at strategic and operational levels), and those involved in information collection and analysis in humanitarian organisations. It is also intended to be read by academics and students of humanitarian action, as a contribution to the larger debate and discussion on this topic.

The specific questions considered in this paper are:

Section 1: Why does evidence matter for humanitarian action?

Section 2: What is 'evidence' in the context of humanitarian action and how do we judge its quality?

Section 3: Does current evidence meet the criteria for quality?

Section 4: Is current evidence used to guide humanitarian operations and policy-making?

Section 5: How can the quality and use of evidence in humanitarian action be improved?

The paper is based on a literature review, augmented by interviews, and on the content of presentations made at the ALNAP 28th Annual Meeting , 'Evidence and Knowledge in Humanitarian Action' (Washington DC, March 2013). Further details of the methodology are given in Annex 1.

Box 1. Definitions

For the purposes of this paper, we adopt the following working definitions.

Key terms

Evidence: information that helps to substantiate or prove/disprove the truth of a specific proposition.

Information: any data that may inform understanding or belief, presented in a context that gives the data meaning. Information may be true or false. Information only becomes evidence when it is related to a specific proposition.

Knowledge: 'justified true belief'. In the context of this paper, knowledge is understood to derive either from direct observation or from a body of evidence to inform a true understanding of a particular topic.

Research terms

Bias: any form of systematic (non-random) error.

Hypothesis: a proposition, supported by a limited amount of evidence, offered as a potential explanation for a phenomenon. A hypothesis is often the starting point for further investigation: further evidence can be collected to further support (or disprove) the hypothesis.

Qualitative: (of research, analysis, data) based on narrative, rather than numbers. Qualitative research tends to relate to human behaviours and motivations.

Quantitative: (of research, analysis, data) based on numbers rather than narrative. Quantitative research is based on (statistical) analysis of a dataset.

Reliability: relates to the consistency of the results of the same experiment when repeated: it is a measure of the degree to which the same experiment will produce the same result under the same conditions on different occasions.

Scientific method: a method or procedure that has characterised natural science since the 17th century, consisting in systematic observation, measurement and experiment, and the formulation, testing, and modification of hypotheses (Oxford English Dictionary).

Validity: a term with a variety of related meanings, all of which are concerned with the degree to which an experimental argument or conclusion is 'well grounded'. In the social sciences, and in qualitative research in general, validity generally relates to the strength of the relationship between a conclusion or analysis and the observed data (LeCompte and Goetz, 1982) or more broadly, to the strength of the relationship between a statement and the social phenomenon to which it refers (Hammersley, 1990). In experimental science, validity has a series of specific meanings. Internal validity is a measure of the strength of cause and effect relationships observed through statistical tests: the degree to which a particular effect can be attributed to a particular cause. External validity, on the other hand, concerns the extent to which a cause and effect relationship identified in a particular situation can be generalised to other situations.

2

What is 'evidence' in the context of humanitarian action and how do we judge its quality?

2.1 Background: terms, concepts and challenges

This paper is based on a few core concepts that are important to clarify from the outset. The meaning of **knowledge** has been much debated by philosophers down the ages, and that debate is at the heart of the philosophical subject of epistemology (which concerns the nature and scope of knowledge). For our purposes, we take it to mean a 'justified true belief' – in other words, a belief that is grounded in some form of 'fact'. Knowledge differs from information, which we take to mean any data that may inform understanding or belief, presented in a context that gives it meaning.

Within the field of epistemology, the empirical school of thought has particular relevance to our discussion. For empiricists, knowledge comes from sense-based experience ('the evidence of one's eyes'). Empiricism is opposed to idealism, tradition, authority: the idea that something is right because the theory says so, or because that is how we have always done it, or because that is what the experts tell us to do. It is this attitude that has characterised much of Western scientific thought for the past 400 years, and which underlies the scientific method: forming hypotheses, and subsequently testing – and as required modifying – these hypotheses on the basis of experiment and observation. In the social sciences, empiricism is often recast as 'positivism', whereby propositions that cannot be tested using observation and experience, unless they are true by definition or logical inference, are meaningless.

Social scientific approaches are not, however, limited to the empirical or positivist approaches embodied in the mainstream Western traditions. Since the mid-20th century, positivism has been challenged by competing viewpoints in the social sciences. Broadly speaking, this has involved a challenge to the existence of 'absolute' truth and knowledge separate from the observer, and with it the idea of strict objectivity. It has also involved a challenge to 'linear' models of cause and effect. Emphasis is placed instead on the relative and subjective, on people's perceptions and experiences, and on more complex – and socially and politically informed – explanations of behaviours and outcomes. This approach tends to privilege qualitative methods over quantitative approaches with their strict focus on what is measurable. Contemporary approaches to social and cultural anthropology largely fit within this paradigm, and it informs many other areas of social scientific and policy thinking.

The idea of 'evidence', as used in this paper, is closely connected to empiricism and the scientific method. Evidence is information which helps to demonstrate the truth or falsehood of a given hypothesis or proposition. In many cases, these propositions relate to the existence or absence of a condition (malnutrition, or a disease epidemic) which can be verified through repeated and systematic observation, conducted according to proven methodologies. In some cases, however, the propositions will relate to the behavior or beliefs of (often diverse) groups of people, and here – while broadly following the approach of developing and testing propositions – we will need to recognise that a single, objective 'reality' does not exist, and that our propositions instead relate to multiple, and perhaps conflicting, ideas. This in turn requires us to investigate the value of a variety of different types of information as evidence, and to use a variety of non-experimental methodologies to consider the value of this information in supporting the proposition.

"Evidence is information which helps to demonstrate the truth or falsehood of a given hypothesis or proposition."

If humanitarian evidence differs in type, it also differs in quality. Nutritional data gathered using a valid method of nutritional assessment is better evidence for the existence of a nutritional crisis than an anecdotal report of large numbers of wasted children (although both are evidence, of a sort). We should be careful, though, not to assume that differences in evidential quality are simply based on the nature of the evidence (qualitative or quantitative): a structured questionnaire, administered to a probabilistic sample of the population, may well tell us less about ways to improve a livelihoods programme than a series of well-conducted semi-structured interviews. A recent publication from DFID, *Assessing the strength of evidence* 'avoids constructing a hierarchy of research designs and methods…[and] recognises that different designs are more or less appropriate to different contexts, and different research questions' (DFID, 2013: 7). The challenge – which we address in Section 2.3 – is to identify criteria for the evidential quality of a variety of different types of information. In many cases, we will be able to generate stronger evidence by using 'mixed method' approaches, using 'a wide spectrum of evidence which uses, and triangulates several research designs and methods' (ibid: 8).

Finally, we should note that the quality of the evidence that we will be prepared to accept will also differ, depending largely on the circumstances in which we find ourselves. Sometimes, anecdotal evidence may be all we have to go on, and it may be enough to trigger action of certain kinds; though the action will usually be to obtain more rigorous evidence on which to base further decisions about action.

IN BRIEF:

- In this paper, we define evidence as information that informs a specific proposition.
- This understanding of evidence draws heavily on an empiricist, and broadly 'scientific' understanding of knowledge.
- However, the nature of humanitarian action means that we are often dealing with subjective and socially constructed 'realities', and as a result should be open to considering many different types of information as evidence.
- Evidence differs in quality: this difference is not, however, related to whether the evidence is quantitative or qualitative in nature.
- The threshold for use of evidence will depend on circumstances: in some circumstances we should be prepared to accept lower quality evidence.

2.2 Evidence for what? Humanitarian propositions and evidence requirements

As described in Section 2.1, evidence properly understood is evidence *for* something; specifically, it is information or analysis that goes to support a particular proposition or claim. Since the concept of evidence is linked so closely to the idea of propositions, we need to consider the nature of these propositions in the humanitarian context, and what might be required to demonstrate their truth or falsehood. We suggest it is possible to identify two linked sets of propositions that are critical to the decision-making processes of actors in the formal humanitarian system:

- Type A: propositions about the existence of humanitarian needs resulting from a crisis, and;
- Type B: propositions about 'what works' in addressing these needs.

Within each of these two 'sets' there will be various specific propositions, and each of these will require evidence (often of different types) to support it.

Propositions about the existence of need

The first set of propositions (Type A) deal with the existence of needs that the international community should address. Within this set, the key specific propositions are:

- a situation exists that is 'critical' for those affected, *or* a situation exists that, while not yet critical, will become so if not addressed
- the situation requires intervention by external actors.

These propositions are concerned chiefly with proving that a humanitarian intervention is required. Evidence to support such propositions generally comes from early warning and assessments and is often based on measurement of certain indicators to describe the situation in relation to key aspects of people's lives and livelihoods. In many cases, the results of these measurements are compared with accepted crisis thresholds, to show that the qualifying conditions for a current or impending 'crisis' situation have been met.

Most humanitarian crises – at least as seen through the eyes of professional humanitarians – are made up of sub-crises that are sector-specific: for example, the aftermath of catastrophic floods may see crises of public health, food security and shelter. Many of these sectors have their own indicators and thresholds, although in some cases responses will be based less on the degree to which symptoms have reached explicit thresholds, and more on the judgement of the decision-maker. The choice of threshold, and the lack of thresholds in some cases, raises questions about what constitutes a crisis and what evidence is needed to demonstrate it.

“The challenge of identifying 'a crisis' is particularly acute when the proposition deals with the potential for a future emergency, rather than with a description of the current situation.”

The challenge of identifying 'a crisis' is particularly acute when the proposition deals with the potential for a future emergency, rather than with a description of the current situation. In this case, the evidence needs to show that the situation is developing in such a way that emergency conditions will exist in the future. In some cases, this can be done by a comparison of information about the current situation

with pre-determined 'triggers' for early action: however, in many cases these triggers do not exist –indeed, there are legitimate arguments over whether generic triggers, which are independent of specific contexts, are realistic.

Ideally, Type A propositions also call for evidence to show that the crisis cannot be addressed effectively unless the international community becomes involved. In practice, however, this evidence is not always requested before an intervention is launched.

Propositions about the effectiveness of response

The second set of propositions (Type B) relates to how international actors should respond to a current or impending crisis. Within this group, the key specific propositions are:

- an intervention of a specified type will be (or was) effective in preventing or mitigating the effects of the crisis (or any such crisis) in defined ways
- the proposed intervention is (or was) the most appropriate available in the context, taking into account likely effectiveness, cost efficiency, relevance as perceived by local people, etc. (the 'best option' intervention)
- the intervention can be delivered on the basis proposed, meeting agreed minimum standards – in other words, it is feasible.

These propositions lie at the core of the choice of response when deciding how to intervene in a specific situation. The first compares the effectiveness of an intervention with doing nothing. The second depends on demonstrating why the proposed intervention is the best choice among the options available and is of growing importance to humanitarian agencies as they gauge the relative worth of various response methods. In the past, many may have used a fairly limited range of 'default responses', but these are now being challenged by advances in learning about alternative options. For example, the increasing documentation of lessons about alternatives to standard approaches to food and livelihood insecurity (e.g. cash distribution instead of food aid) is putting growing pressure on agencies to justify their choice of response (Maxwell et al., 2012).

The third proposition concerns the criteria used to determine which intervention is 'best': it may, for example, be cost-efficiency; cost-effectiveness; acceptance by the population; the ability to reach all people in need; the degree to which the intervention articulates with other interventions to meet all needs in a coordinated way; or some combination of these.[2] Evidence related to these different criteria will require different information sets, and different methods for collection and analysis.

Variants of these propositions – considering past actions – are also central to evaluation when considering whether a response was successful. In general, this requires evidence of the results of the intervention, and presents the often knotty challenge of proving that this specific intervention led to those specific results.

2 The generally agreed criteria for determining the success of an intervention are those proposed in the OECD DAC evaluation criteria (see Beck, 2006).

The evidence required to address humanitarian propositions

Both Type A and Type B propositions tend to require two broad types of evidence: specific and general. Specific evidence tells us what is happening at a particular time in a particular place. General evidence is gained by studying a number of similar emergencies, in order to be able to say something sensible about 'how famines develop' or 'how public health interventions work'. By considering both sets of evidence together, we can situate what is happening in a broader context, and can say with some certainty – for example – whether or not these conditions constitute a crisis when compared to situations elsewhere, or whether this is the sort of situation where a particular type of response is often successful.

In some cases, general evidence – about how famines develop, or how shelter can best be provided to refugees – is codified into thresholds, triggers, or standard procedures, all of which are based on the idea that what has happened in other contexts can be extrapolated with confidence to this particular context. These triggers and procedures can greatly increase the speed of response, and can, importantly, create consensus around whether a response is required, and what form it should take. However, because every emergency situation is different, we should treat most general evidence – and its resulting thresholds, triggers and procedures – with some scepticism, and always ensure that it is relevant to the specific situation of concern.

It is also worth noting that evidence around both sets of propositions seldom comes in the form of direct measurements. More often, our evidence takes the form of indicators: pieces of information that do not describe or measure the issue of concern directly, but that point to, or 'indicate' the issue. So prices, nutritional status or crop yields, while not providing a direct measurement of food insecurity, may all point to food insecurity. In general, indicators are used on the grounds that they are (relatively) simple to measure and have a reasonably reliable and consistent relationship with the object that we wish to know about. However, they are rarely conclusive, and often have to be used in combination. A key evidential concern in humanitarian action, particularly for diagnostic (Type A) propositions, is the selection of indicators: where the wrong indicators are selected, it is possible to have perfectly good information that is of no relevance whatsoever to the specific proposition.

> **"Where the wrong indicators are selected, it is possible to have perfectly good information that is of no relevance whatsoever to the specific proposition."**

Before we move on, it is important to make one caveat: the two sets of propositions that are considered in this paper – and the evidence that is used to support them – arise from the operational requirements of international humanitarian organisations. While they form the focus of this paper, these two sets of propositions do not necessarily encompass, or admit, all of the knowledge about a crisis that exists. The paper is written predominantly from the point of view of international humanitarian actors, but we should remember that other groups involved in the emergency – and particularly the people affected most directly – may have other questions, and may well require other forms of evidence and knowledge to answer them.

IN BRIEF:

- Most evidence collected by international humanitarian organisations is designed to address propositions about the existence of need; *or* propositions about the best way to respond to this need.

- Evidence related to these propositions comes in many different forms, and relates to many different topics.
- Humanitarian decisions are often made by comparing evidence from a specific situation with 'general' evidence on how similar situations developed or were addressed in the past, but caution is needed: 'general' information from previous responses may not always be relevant.
- Much of the information used by humanitarians comes in the form of indicators, but care is needed in their selection to ensure that they have a strong and consistent relationship with the condition being measured.

2.3 Criteria to assess the quality of humanitarian evidence

We run into a number of problems when trying to substantiate the propositions outlined in Section 2.2. First, we need to satisfy ourselves that the propositions are true, based on the best available evidence. But how do we know if we can trust the evidence? Second, how do we judge whether the evidence actually supports the proposition in question? In this section, we consider some of the relevant issues and propose six criteria for 'good evidence'.

Evidence in the contexts we are considering comes both as: data, whether quantitative or qualitative, direct or indirect; and as the result of analysis of such data, such as the conclusions of assessments, evaluations or studies. When considering whether information can be used as 'evidence' – that is, whether it advances or disproves a proposition – we need to be aware that it is not only the quality of the data that matters, but also the quality of the analysis. It is quite possible for good data to be badly analysed, and to lead to the wrong conclusions. Before saying that information counts as evidence, both the data, and the methods used to analyse this data, should be measured against specific quality criteria.

As suggested above, evidence of different kinds can be described in terms of a number of key attributes, including truth, credibility, accuracy, reliability, and validity. Unfortunately, these terms are used in different ways across different disciplines and sectors, making it difficult to generalise about their meaning. To complicate matters further, some of these terms are used to describe both the quality of data and the methods by which data are collected and analysed.

Here we have attempted to summarise the six main criteria that might be used to assess the quality of evidence in humanitarian contexts.

i. **Accuracy:** *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.

REPRESENTATIVENESS

ii. **Representativeness of the evidence:** *the degree to which the evidence accurately represents the condition of the larger group of interest.* Information that passes the *accuracy* test to, for example, reflect the situation in one village or the views of one population group may not represent all the villages or population groups in the area. This is a question of the degree to which a sample represents the whole population.

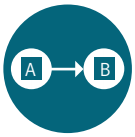
RELEVANCE

iii. **Relevance:** *the degree to which a piece of information relates to the proposition that it is intended to prove or disprove.* A piece of information can be accurate and representative, but may still be irrelevant to the argument or the proposition being made. The issue of *relevance* often relates to indicators. These are not direct measurements of a condition, but they measure something that points to, or indicates that condition. Anthropometric measurements, provide reliable indications of, for example, nutritional status: there is a strong relationship between the indicator and the condition, so these measurements are relevant as evidence to prove the proposition that 'malnutrition exists in this area'. However, the same measurements are less effective as an indicator for food scarcity, as the relationship between arm circumference and lack of food is less strong (the malnutrition could be caused by illness or poor feeding practices). As a result, MUAC measurements are less relevant as evidence to prove the proposition that 'food aid is required'. More generally, the test of *relevance* should be applied to any information which is provided as evidence in support of a particular argument or proposition.

GENERALISABILITY

iv. **Generalisability of conclusions:** *the degree to which evidence from a specific situation can be generalised beyond that response to other situations (and particularly to situations in other countries at other times)*, and so used as more general evidence of how a situation will unfold, or of the best type of response. The issue of *generalisability* is of particular concern to policy-makers, who attempt to create policies with global application from evidence generated in specific contexts. It is also at the heart of discussion about common global indicators and thresholds for disasters.

ATTRIBUTION



v. **Attribution of analysis:** *whether analysis demonstrates a clear and unambiguous causal linkage between two conditions or events.* This is central to the evidential quality of evaluation: where evaluators identify certain changes in circumstances that occurred after an intervention, how confident can they be that these changes were the result of that specific intervention, and not something else?

CLARITY AROUND CONTEXT AND METHODS

vi. **Clarity around context and methods:** *the degree to which it is clear why, how, and for whom evidence has been collected.* 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'.

This classification of criteria is by no means authoritative, and there are a number of other approaches to testing the quality of evidence. In a 2005 paper on the use of evidence for policy-making, Louise Shaxson – building on Spencer *et al.* (2003) – suggests five components that, taken together, define the 'robustness' or strength of evidence in policy terms: credibility, reliability, objectivity (lack of bias), rootedness,[3] and *generalisability* (Shaxson, 2005). The group British Overseas NGOs for Development (BOND) proposes five principles for evaluating the strength of evidence produced by (development and humanitarian) NGOs: whether the perspectives of people living in poverty are included in the evidence; whether the evidence is generated through methods that are justifiable given the nature of the assessment; the degree to which data are triangulated; the degree to which evidence explores how change happens; and whether the evidence discloses the details of the data sources and methods used, the results achieved, and any limitations in the data or conclusions.

John Gerring, in proposing a unified framework for social scientific methodologies, suggests some criteria that apply to all social science arguments (propositions), including truth, precision, generality, coherence, commensurability and relevance. But as he observes, 'Distressingly, the vocabulary associated with the subject of methodology is ridden with ambiguity. Key terms… mean different things in different research traditions and different research contexts.' (Gerring, 2011: 16). So we should not be surprised that a single list of agreed criteria is hard to produce. However, we hope that the six criteria given here, which reflect the general consensus of the humanitarian practitioners and academics at the 28th ALNAP Annual Meeting,[4] provide a useful tool for humanitarian actors when considering the strength of the evidence that they use.

IN BRIEF:

- Different sets of criteria exist to judge the quality of evidence, although they often refer to a specific type of evidence (qualitative/quantitative) and are not, therefore, generally applicable to all humanitarian evidence.
- This report uses six criteria to judge the quality of evidence generated and used in humanitarian action: *accuracy; representativeness; relevance; generalisability; attribution; and clarity around context and methods.*

3 'Rootedness' is about whether the question being addressed by the evidence truly represents the fullness of the issue concerned or whether there are other aspects that could and should be explored. The idea relates to that of *clarity around context and methods*, outlined above.

4 An initial list, broadly encompassing the first five criteria, was presented to the meeting and discussed in plenary. While there was no objection to the criteria themselves, there was a general sense that they did not address who the evidence is for, and for what purpose it is collected. This concern has been captured in the sixth criterion. Delegates at the meeting also suggested that 'timeliness' be included as a criterion. This has not been included in this paper, as it seems, to the authors, to be more related to the utility of the evidence than the intrinsic quality of the evidence itself.

3

Does humanitarian evidence currently meet the criteria for quality?

3.1 The challenges to generating evidence in humanitarian situations

There are a number of serious challenges to the generation of good evidence in the humanitarian sector, many of which stem from the context within which humanitarian actors operate. International actors often become involved in responses where state institutions are fairly weak. As a result, secondary data are often very limited in the areas where humanitarians work: even basic population data can be hard to come by (Box 2).

Where good secondary data are available, the nature of humanitarian emergencies – often involving high levels of mortality and significant population movements – means that information can soon be out of date. Population figures, malnutrition rates, information on shelter: all can change very quickly, creating a requirement to collect information continuously as the situation evolves.

Unfortunately, the circumstances that make information collection so important are precisely those that make it extremely difficult to do. Many humanitarian responses occur in situations where physical access is severely restricted, limiting possibilities for data collection. At the same time, the distribution of large quantities of goods and services in contexts where there is immense need will, inevitably, have political implications; particularly in 'complex emergencies' where there are already political and factional rifts between different groups within an area. Under these circumstances, humanitarians cannot assume that information – and particularly information from key informants – is objective or accurate. Nor can they automatically assume that they themselves are free of bias.

“Unfortunately, the circumstances that make information collection so important are precisely those that make it extremely difficult to do.”

Box 2. The question of numbers

Some of the most basic numbers that matter for humanitarian propositions are notoriously arbitrary. In particular, population estimates and figures given for 'numbers affected' or 'numbers of beneficiaries reached' tend to suffer from a high degree of uncertainty and a lack of definitional clarity. This threatens to undermine the credibility of the propositions that include these numbers.

Frequently, the root of this problem lies in a lack of certainty about baseline population (denominator) figures. The best source of baseline population data is normally the census, but census data may not a available at local level, and where the census is designed and used as a political tool, the numbers it provides are not reliable. Even where this is not the case, census figures are

often out of date, and while there are accepted methods to create population projections, these rely on critical assumptions that may be difficult to make with any certainty.

In the absence of census data, humanitarians have resorted to a variety of approaches: quick counts – either on the ground or using satellite imagery – of houses/buildings; headcounts; listing and profiling; flow monitoring; and randomised household surveys. These approaches are, however, time consuming and vary in accuracy. Frustratingly, they tend to be more successful where census data already exist.

One unfortunate consequence of this lack of baseline data is that the diversity within a society is often hidden. Gender and age differences, in particular, which are an important element of most censuses, are seldom collected among the social groups assessed and data disaggregated by age and sex remains the exception rather than the norm.

Sources: Demographic Assessment Techniques in Complex Humanitarian Emergencies: Summary of a Workshop (2002) Committee on Population, National Academies Press, US; 'Sex and Age Matter', Tufts University (2011); State of the Humanitarian System, ALNAP (2012); Guidelines on Data Issues in Humanitarian Crisis Situations, UNFPA (2010); Technical Brief: Estimation of Affected Population Figures (ACAPS, 2012).

Humanitarian operations are also marked by the accentuated power differences that exist between outsiders (who control resources and understand the humanitarian system) and the people affected by the emergency. The latter may be in great need. At the same time, they tend to have only limited understanding of how the system works and very little time to complete questionnaires or participate in interviews. Not only does this create significant practical challenges to the collection of primary data, it also leads to worrying ethical challenges: to what degree is it legitimate to ask people to provide information when they might not directly benefit themselves? To what degree can affected people be said to be giving informed consent over the use and release of information in a system they don't understand, which they may hope will provide them with material support? As Bradt argues, 'Data-gathering and consequent humanitarian interventions are invasive procedures with unintended consequences. Good intentions do not excuse bad outcomes' (Bradt, 2009: 15).

A further ethical challenge arises from the fact that evidence generation requires funds (and time) which, in most cases, come from budgets that could be spent on direct support to affected people, creating a difficult balance between evidential rigour and the need to respond.

Finally, humanitarian operations are often characterised by the need to obtain information in very short time frames. It is important not to overstate this point: the states where most humanitarian expenditure has occurred over the past five years are all affected by chronic conflict (ALNAP, 2012), and much of the humanitarian work that occurs in these places takes the form of long-running programmes where there is time to collect and analyse information (Macrae, 2013). However, in rapid-onset emergencies, such as the Haiti earthquake in 2010 or Typhoon Yolanda (Haiyan) in 2013, time is tight, and this affects the ability of humanitarian actors to collect enough good evidence to support effective action in the early phases of a response.

All of these constraints emerge, primarily, from the context of humanitarian action at the individual project or programme level. However, problems also exist at the global or systemic level – not least because of the fact that humanitarian operations in Afghanistan and Zimbabwe differ greatly in both context and nature. Establishing evidence of 'what works' under these circumstances can be extremely challenging in a sector where social and cultural variables are crucial for the success of so many operations.

The rest of this section considers the ways in which organisations are addressing these challenges, looking at how evidence and knowledge are generated in practice in the humanitarian sector, and at related questions about the quality of this evidence. Our focus is on current practice in the sector as it relates to the 'standard' elements of humanitarian programming: the evidence generating processes of early warning and surveillance systems, needs assessment and monitoring processes, and project or programme evaluations. We consider the nature of evidence generated by each, particularly as judged against the 'quality' criteria set out in Section 2 above.

"Humanitarian operations in Afghanistan and Zimbabwe differ greatly in both context and nature. Establishing evidence of 'what works' under these circumstances can be extremely challenging."

IN BRIEF:

- There is generally an urgent need for information in emergencies and crises.
- But only limited baseline data are available in many contexts where international assistance is required.
- Collecting information in emergencies can be extremely difficult, with constraints to access; information that may be used as a political tool; and important ethical considerations to be taken into account.

3.2 Evidence from early warning systems

Early warning systems have been described as 'combinations of tools and processes embedded within institutional structures… [and] composed of four elements: knowledge of the risk, a technical monitoring and warning service, dissemination of meaningful warnings to at-risk people, and public awareness and preparedness to act…'.[5] Not every early warning process is part of such a joined-up system. Some, notably some famine early warning processes, provide independent data collection and analysis without being prescriptive about action. Others are less formal and more community-based.

Early warning systems combine new data about a developing situation with historical knowledge to predict the likely outcome for a given area over a given time frame. While the data concerned are context-specific, the analysis will draw on both knowledge of context (e.g. livelihood types) and knowledge of previous occurrences in this or other contexts to make situation-specific predictions. That analysis is often trend-based: in other words, its force depends on being able to establish a convincing case for an emergent trend that – if left unchecked – will lead to catastrophic outcomes (proposition Type A above).

5 Source AlertNet: www.trust.org/alertnet/news/early-warning-of-disasters-facts-and-figures

Famine early warning is the best known example of this type. The case to be made is usually complex, combining data of different kinds, such as rainfall patterns, harvest yields, food prices, terms of trade, household income and malnutrition levels. Other cases may be simpler, such as early warning that tracks the likely path of an approaching cyclone or the likely onset of a flood, or even of an impending earthquake or volcanic eruption, though these are harder to predict.[6] Early warning can buy time to take preventive, preparatory or evasive action.

Early warning is one of the areas of humanitarian practice that has advanced most in the past 20 years. From famine early warning systems like FEWS Net and FAO's Integrated Phase Classification (IPC) system in Africa, to cyclone-tracking systems in Asia and the Caribbean, advances in technology combined with coordinated national and cross-border systems and effective community mobilisation have helped to reduce vulnerability. Combined with other elements of preparedness (e.g. flood shelters in Bangladesh, food-aid pipelines in the Horn of Africa, community response mechanisms in India and Central America), these systems have saved many lives.

The evidence from most established early warning systems generally appears to score well for *accuracy* when judged against the criteria for strength of evidence. But many systems struggle with issues of *relevance* – linking the data that has been collected to the proposition that a crisis is coming that will have a major impact on lives and livelihoods.

"Complexity makes it hard to tell decision-makers a clear and convincing story."

In rapid onset crises where (for example) cyclones and floods can be closely tracked, it is increasingly possible to say with a fair degree of certainty if the circumstances are likely to lead to catastrophic human impact and which areas will be affected. However, in cases of slow onset disasters, such as drought-related food-security crises, it may be much more difficult. In these crises, multiple factors combine to determine the impact on people and their livelihoods. This makes the choice and weighting of indicators extremely difficult: are crop yields or cereal prices better predictors of future food insecurity? And how are they related? The complex interplay of factors also means that negative indicators may not always lead to bad situations, and unforeseen factors often come into play. This complexity also makes it hard to tell decision-makers a clear and convincing story.

The late response to food and livelihood insecurity in the Horn of Africa in 2010/11 highlighted these difficulties, and focussed attention on the several early warning systems operating in the region. In general, observers concluded that the systems 'provided accurate and timely information that enabled those in positions of power to plan and respond' (Hillier and Dempsey, 2012: 14; See also Majid, 2011; Slim, 2012). A joint Oxfam GB/Save the Children report noted that 'FEWSNET and FSNWG reports were graded as 'very good' to 'excellent' in terms of their *accuracy* in predicting the severity and onset of the crisis'[7] (Hillier and Dempsey, 2012: 14).

However, even in these effective systems, patchy data led some people to question the *accuracy* and *representativeness* of early warning reports (Levine et al., 2011).

6 UNISDR, 2006.

7 To a degree this may be because Early Warning deals with uncertainty; with situations which *may* exist in the future, rather than situations that can be measured now. This uncertainty leads decision-makers to demand a higher standard of evidence, and so of certainty, before they will act.

This challenge was compounded by the fact that the methods used to analyse the multiple, inter-related sets of information were not transparent – there was a lack of *clarity around context and methods* (ibid). The authors of the report *System Failure: Revisiting the problem of timely response to crises in the Horn of Africa* recognise that one way to address this problem of analysis and interpretation would be to identify 'the perfect indicator that cannot be manipulated and will tell – in every District – which intervention is needed' (Levine et al., 2011: 11). However, in solving one set of problems around evidence, this would create another, as no single indicator could capture livelihood insecurity accurately in all contexts (in terms of the typology outlined above, no single indicator would be *relevant* in all cases: a problem of *generalisability*).

Paradoxically, debates over the best methodologies and indicators to use for early warning – debates that are intended to improve the quality of evidence – appear to have made decision-makers more, rather than less, sceptical about the value of early warning systems. We will return to the theme of their use of evidence – and particularly the use of evidence from early warning systems – in Section 4.

IN BRIEF:

- Early warning is an area that has seen significant investment and improvement in recent years.
- The key evidential challenge in early warning is selecting relevant indicators that can forecast conditions reliably, particularly where early warning addresses crises (like famines) that are the result of many factors interacting over time.
- To succeed, early warning systems need to be clear about their methods – particularly for analysing and weighing many different types of evidence.

3.3 Evidence from needs assessment

'Needs assessment', as we describe it here, encompasses a wide range of activities that aim to identify whether external assistance is required and, if so, outline the type, quantity, and timing of that assistance. In the context of this paper, then, most needs assessments aim to provide evidence related to propositions around the existence of a crisis (Type A propositions) and to a lesser degree, around the best way to address a crisis (Type B).

Needs assessments are affected by the generic constraints to evidence collection identified in Section 3.1: lack of basic data; difficulties of access; and shortage of time. However, they also face specific challenges that spring from the very nature of the task. Above all, emergencies tend to generate a wide variety of different types of need. Some of these – and particularly those related to health and nutrition – can be predicted, and can be captured by measurement of objectively verifiable indicators, using standard methods and protocols. However, other needs, such as food insecurity, are conceptual categories, which tend to be measured by various combinations of indicators. As we noted in the section on early warning, these indicators will often differ from one place to another and, as a result, there is little common agreement on which should be used and how they should be weighted. There are other needs that have a strong subjective component: feelings of threat

or insecurity, for example, point to needs that merit attention but that are hard to capture through quantitative methods of measurement (Mahmood et al., 2010).

As a result, needs assessments require – ideally – individuals or teams skilled in a variety of different disciplines, with the ability to collect, analyse and reconcile a variety of different types of evidence. In practice, this requirement to synthesise evidence and to 'fill in' any evidence gaps leaves many needs assessments heavily reliant on the contextual knowledge of the individuals involved in the response (see Poole and Primrose, 2010; Assessment Capacities Project (ACAPS), 2012; ICRC and IFRC, 2008).

Contextual knowledge also matters for determining the scale of assessment required. In some cases – where repeated disasters of the same type have had very similar consequences – contextual knowledge may allow humanitarians to avoid or minimise assessment, and move straight to programming. It can also help identify the best methods and approaches required to conduct an assessment in any given place. The availability of information and of skilled staff, and the ability to access the emergency area, will all differ from one situation to another. As a result, 'one size does not fit all…there is not a single standard [assessment] package that can be rolled out at each disaster. Rather best practices about good assessment approaches need to be adapted to meet information needs specific to each crisis' (Assessment Capacities Project (ACAPS), 2012: 6).

The danger here is that without a single standardised assessment approach, each agency in each emergency will create their own assessment methodology for their own separate assessments. A certain amount of duplication in assessment – particularly where this duplication is planned and coordinated – is probably a good thing: multiple assessments can increase the evidential quality of information by allowing results to be compared. However, the situation in the past has tended to be one where 'most agencies had their own non-standardised survey forms…that often produced conflicting or repetitive results' (Darcy et al., 2013: 24). Obviously, this is not an efficient use of scarce resources. Different agencies collect similar information, while failing to fill obvious information gaps[8]; people in disaster-affected communities spend precious time answering the same questions on multiple occasions. In addition, the resulting reports are often only used by the agency that created them because they are not shared (Mills, 2005); are not shared in a timely manner (Darcy and Garfield, 2011); are not relevant to other stakeholders (Bradt, 2009), or are not in a format that is useful to other actors (Poole and Primrose, 2010).

“The answer seems to lie in increased coordination between assessments, and in the creation and acceptance of common standards and methodologies for data collection.”

The answer seems to lie in increased coordination between assessments, and in the creation and acceptance of common standards and methodologies for data collection (DFID 2012: 31; Poole and Primrose, 2010: 1) that, while they can be adapted for specific contexts, use broadly similar language, indicators and approaches. Fortunately, recent years have seen greater efforts to increase coordination and to develop broad standards, supported by key actors such as IFRC/ICRC, the IASC's Needs Assessment Task Force (NATF), and the Assessment Capacities Project (ACAPS).

8 Such as the collection of data disaggregated by age and gender, see (Mazurana et al., 2011.

In terms of coordination, there appears to be a general move towards the acceptance of joint or coordinated assessments, which, as Garfield et al. have noted, have the potential to increase efficiency and to provide a stronger basis for coordinated planning and action (albeit with potential costs in terms of slowness and expense) (Garfield et al., 2011). In one example, quoted in the *State of the Humanitarian System*, an evaluation found that joint needs assessment conducted by agencies that are part of the Emergency Capacity Building (ECB) project in West Java and Sumatra enabled a response with greater geographic coverage and greater sectoral spread (Wilson et al., 2010).

Meanwhile, a review of the guidance produced by various 'key players' in assessment, and of recent publicly available assessment reports,[9] suggests a growing methodological consensus around approaches: particularly approaches to multi-sectoral early-phase assessments (those assessments conducted in the first three to four weeks after a disaster). The emerging consensus seems to be for assessments based on a review of secondary data – where possible – augmented by information collected from key informants and focus groups of affected people using a questionnaire format. In most cases, the communities to be visited for interviewing are selected according to a purposive, rather than a random sample. While the ICRC/IFRC guidance already offers a suggested questionnaire format (ICRC and IFRC, 2008), the NATF's *Multi-Cluster/Sector Initial Rapid Assessment (MIRA)* guidelines are still provisional (IASC, 2012a) and in the process of being improved (Joyce Luma, personal communication, April 2013). ACAPs, meanwhile, does not provide a questionnaire format, but may consider doing so in the future (Richard Garfield, personal communication, April 2013). In all cases, these formats are not prescriptive, but aim to be 'a starting point [with the expectation that] teams will add questions on' (Joyce Luma, personal communication, April 2013).

Considering the current state of humanitarian needs assessment against the criteria for evidential quality proposed in Section 2.3, it is perhaps inevitable that needs assessments – which will form the basis of distributions of limited amounts of scarce and much needed resources – should encounter problems of *accuracy*. They provide obvious incentives for people to exaggerate the seriousness of a situation to increase the level of support provided. At the same time, governments may underestimate the extent of needs to deny that a crisis is occurring.

In general, the needs assessment guidance suggests that this *accuracy* challenge is best addressed by triangulating key informant or focus group information against the results of other groups or informants, against secondary sources, or, in some cases, against 'constants' such as required calorific intake (see Seaman et al., 2000). 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

9 Reports reviewed were: The Philippines Multi-Cluster Needs assessment for Tropical storm Washi; Multi-Sector Initial Rapid Assessment, Pakistan Floods, 2012; Evaluation Initiale Rapide Multi-Cluster sur les inondations au Moyen Chari, Tandjilé, Mayo Kebbi Est – Tchad (all of which were available at Humanitarian Response. info: http://assessments.humanitarianresponse.info/mira-reports. A fourth report, available on the same page, was not considered) and Joint Rapid Assessment of the Northern Governorates of Yemen; Joint Rapid Assessment of Northern Syria – Aleppo City; and Rapid Initial Needs Assessment – Haiti (all of which were found as a result using the search term 'assessment report' on the ACAPS website. Two further reports identified by the search were not considered).

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, 2013). A 2011 assessment in Yemen took a similar approach, (ACAPS, 2011). However, other reports reviewed for this paper did not specify the methods used to establish the veracity of key informant and focus groups responses.

> **“It is important to make assessments publicly available, to allow findings to be considered and challenged by authorities outside the country.”**

A second and equally important challenge to the *accuracy* of assessments comes from the reliance that is often placed on expert assessors and analysts to ‘fill the gaps’ or synthesise various information sources. As a result, assessment findings can be biased – consciously or, more often, unconsciously – by the biases of aid professionals on the ground. Again, triangulation has a role to play here: Red Cross/Crescent guidance suggests establishing diverse assessment teams that cross-check each other’s work, while the IASC guidance stresses the importance of making any biases and knowledge gaps explicit in the assessment report, and explaining how the conclusions were achieved. In this context, it is also important to make assessments publicly available, to allow findings to be considered and challenged by authorities outside the country. While all of the reports considered for this paper addressed the issue of ‘assessor bias’, it was interesting to see how few assessment reports from the past three years are readily accessible on the internet.

Another challenge to the *accuracy* of assessments is the fact that, in emergencies, the situation (and the resulting needs) can change extremely quickly. An assessment may provide a true representation of the situation in one village on the one day that the team visits, but may well be out of date by the time the report is completed. The approach outlined in the provisional IASC guidance aims to ensure continuity of assessment information over time through a phased process, whereby multi-sector rapid assessments are followed by cluster-led, in-depth assessments (IASC, 2012a).

Clusters have, however, made only uneven progress to date in developing guidance around these in-depth assessments, and they are not yet conducted consistently (Richard Garfield, personal communication, April 2013). Nor would a single, further round of sectoral assessments necessarily solve the problem – which seems to be as much conceptual as practical: the humanitarian community still tends to see assessments as ‘one off’ events, rather than as on-going processes, and effective assessment suffers from the same constraints as programme monitoring as a result (see below) – notably a lack of funding and institutional support.

Areas that see regular emergencies, and where one might expect effective and consistent needs assessment approaches to have developed over the past 20 years, are still relatively poorly served (see Poole and Primrose, 2010; Slim, 2012). Even in those (fairly rare) contexts where sequential assessments have been conducted, the use of different questionnaires and assessment methods have made it impossible to compare results and, therefore, to understand changes in needs over time (Garfield et al., 2011). The situation described in the Real-Time Evaluation (RTE) of the Horn of Africa response could stand for most humanitarian operations: although the quality of assessments was good, the system ‘lacked on-going monitoring… [information] was often outdated by the time it was processed and read by the HCT [Humanitarian Country Team]’ (Slim, 2012: 10).

The second element of evidential quality considered in this report is *representativeness*: in this case, the degree to which the information collected for the assessment is representative of the wider group requiring assistance. As suggested above, most needs assessments use purposive sampling to select the communities, households or individuals to be engaged in the assessment. The guidance from the NATF and ACAPS emphasises that, as a result, assessment findings 'cannot represent the whole disaster-affected population' although they can be used to 'understand the most pressing concerns, issues and needs' (IASC, 2012: 21) and in many cases, will form the best available evidence for an intervention.

Random, probabilistic sampling – which allows for the statistically significant extrapolation of results from the sample to the broader population – is much less usual in emergency assessments, although it does occur. Needs assessments in Pakistan and in Haiti have both attempted to use randomised sampling techniques (ACAPS, 2010; UNOCHA and Pakistan NDMA, 2012): the Pakistan survey using geo-referenced population data to establish a sample, the Haiti assessment relying on spatial sampling – assessing geographical areas on the basis of a grid drawn over affected areas, without reference to population density or distribution. Both cases are instructive. In Pakistan, representative sampling was only possible because the Government had fairly current population data, and because enough trained people were available. While this increased the *representativeness* of the results, enumerators were unable to reach 23% of sampled villages. In Haiti, population movement – combined with a lack of population data – meant that while the sample was random, it was not necessarily representative (Garfield et al., 2011). While it would be unwise to dismiss randomised sampling in all contexts, 'statistical methods [for sampling in early phase assessments] are not normally feasible because of time and access constraints' (ICRC and IFRC, 2008), as well as lack of baseline information, funding and staff skills.

As well as identifying a group that can represent the whole population – through either purposive or probabilistic sampling – needs assessments also need to collect data separately for different sub-groups in the population, which will experience different types and levels of need. Effective assessment and understanding of need requires an understanding of the different situations of women, children, the elderly, people with disabilities and those who are socially marginalised (Mazurana et al., 2011; Jacobsen and Furst-Nichols, 2011; Human Rights Watch, 2011; Nakagawa and Shaw, 2004).

In the rare cases where it is feasible and desirable to use statistical methods, the problem can be addressed by stratified sampling (if using random sampling methods). When using purposive sampling techniques, ACAPS suggest that secondary data be used to identify different social and economic groups and different livelihoods areas. On this basis, assessments increase the sample size until the situation for all the different groups has been described (ACAPS, 2012b). However, while some of the needs assessments considered for this review had made obvious attempts to obtain comprehensive data on the specific needs of vulnerable groups – the Yemen assessment, in particular, considers the specific needs of women in each sector – the overall picture is still very similar to that outlined by Mazurana and colleagues: one of 'limited, ad hoc, sporadic use of SADD [Sex and Age Disaggregated Data] in Phase I and II [assessments]' (Mazurana et al., 2011: 3).

Humanitarian assessments can also face problems with the *relevance* of the information they include, and with the indicators they use to identify need. Assessment reports often include information because it is available, and not because it tends to prove or disprove the existence of needs. Similarly, and perhaps more worryingly, assessments have, in the past, used indicators that have had only a very weak or indirect relationship to the situation being assessed. Data on crop yields, for example, are likely to tell us little about food security in areas where the crops in question are not an important source of food for local people. In assessment, 'an understanding of the context is essential in deriving meaning from…indicators, as livelihoods and coping strategies will vary' (Catley et al., 2008).

Perhaps the most serious evidential challenge for assessments arises in the area of *clarity around context and methods*. The basic problem here is one of assumptions. Needs assessments are often perceived to be – and sometimes even present themselves as – objective, comprehensive descriptions of the situation in an area following an emergency event. Under this assumption, needs assessments are seen as answering the question: 'what is the situation here?' with the resulting information assumed to be for use by a broad group of interested stakeholders. The reality, of course, is rather different. Most needs assessments are in fact answering a variation of the question: 'what and how much do we need to provide?' and the information is primarily for use by the actors who collected the information to argue for funding and to plan operations.

This problem of *clarity around context and methods* leads to three separate but related evidential challenges. The first is a tendency to assume (rather than to prove) that a need exists, and that this need cannot be met by the endeavours of local actors: government; civil society organisations; and affected people themselves. The result can be to provide assistance that – at best – is not needed, and at worst, 'engenders passivity…undermines initiative [and]…. feeds dependency and powerlessness' (Anderson et al., 2012: 22).

Assessment questionnaires in the past have tended to be structured around these assumptions (with questions such as: 'how many people lack adequate shelter' and 'how many households in your community face food shortages today?') and have often overlooked the capacities of existing and local structures to respond to the situation. More recent guidance on needs assessment emphasises the need to ask open questions that do not assume a need exists (IASC, 2012b). The guidance also clarifies that the purpose of assessment is not to plan operations, but to determine whether an operation is needed or not (ICRC and IFRC, 2008). These changes seem, to a degree, to have influenced assessment practice: the majority (though not all) of the assessments considered in this review took some account of local capacities to respond to need. Perhaps unsurprisingly, those assessments that were conducted in collaboration with local authorities or organisations (such as those in Pakistan and Chad) appeared to be strongest in this area (OCHA, 2012; UNOCHA and Pakistan NDMA, 2012).

The second challenge relates to the size of the need. The fact that assessments are the basis for funding decisions, and are not neutral descriptions of the post-emergency situation, could provide an incentive for agencies to overstate the true extent of need (to increase funding) or understate the true extent of need (to be 'realistic' about the funding that is available) (Darcy and Hofmann, 2003). One

recent example of this was the failure of needs assessment in Somalia in 2011, which was partly a result of 'pressure to stay low, showing a tendency to shape humanitarian requests to what…donors might give' (Slim, 2012: 10).

Assessments – which outline needs that the State is unable to fulfil – are also intensely political acts, and figures can be 'routinely trimmed down' as a result (Slim, 2012: 14) at the analysis stage (see also UNOCHA and Pakistan NDMA, 2012; Darcy et al., 2013). Here again, the humanitarian system seems to be aware of the need to make assessments more impartial: the establishment of ACAPS was spurred in part by a desire to provide decision-makers with independent assessments. However, there is still some way to go before independent assessments of need are available for all emergencies: for now, 'the system [still] lacks a hard set of numbers on total people in need' (ALNAP, 2012: 42).

The third and final challenge revolves around the understanding of need. Various studies have shown that local perceptions of assistance often differ markedly from those of international actors (see particularly Dijkzeul and Wakenge, 2010). In particular, the understanding and prioritisation of 'needs' for disaster-affected people differs from one context to another (Anderson et al., 2012; Harragin and Chol, 1998). Despite this, most needs assessments are organised 'sectorally', with questions that ask about certain pre-determined types of need. As a result, they provide evidence related to the needs that international actors are best able to address, rather than evidence of need itself.

Similarly, externally imposed definitions of 'vulnerable groups' who are expected to need assistance can differ significantly from the perceptions and understanding of the disaster-affected people themselves. The 2012 ALNAP *State of the Humanitarian System* report found that 'across several contexts and emergency types, evaluators noted a problem with preconceived notions of vulnerability, which led to inappropriate interventions' (ALNAP, 2012: 50).

“The understanding and prioritisation of ‘needs’ for disaster-affected people differs from one context to another.”

There are moves to address this issue. DFID, for example, recently identified a requirement for humanitarians to 'systematically involve beneficiaries in the collection and use of data', noting that this may require new skills and approaches (DFID, 2012: 31), while, on the ground, initiatives such as the Humanitarian Emergency Perceived Needs Scale (HESPER) aim to build a better understanding of need. The development of communications technology also offers greatly enhanced opportunities for humanitarians to access the views of affected people (Box 4).

More work is required, however. While we may have come some way since the damning verdict of the Tsunami Evaluation Coalition (Telford and Cosgrave, 2006), ALNAP's 2012 *State of the Humanitarian System* report found that 'field-based surveys of aid recipients undertaken for this report…revealed that two-thirds of the respondents said that they were dissatisfied or only partly satisfied with the amount and quality of the overall package of assistance that they had received', and that one key reason for this was the failure to understand the local context (ALNAP, 2012; see also Grünewald et al., 2010).

Recognising the local understanding of 'need' in programme design is an important step to ensuring effective programmes (Featherstone, 2013). But the significance of listening to those most directly affected and taking their reality into account has an

impact beyond immediate effectiveness. This is demonstrated by the results of the CDA (Collaborative Learning Project) 'Listening Project' involved conversations with over 6,000 recipients of international aid, who gave a fairly consistent message.

The message is that assessments designed by and for international actors are contributing to a situation where, in emergencies, too much aid comes too fast: a situation which many people see as meeting the institutional needs of international organisations, rather than the needs on the ground. At the same time, the assumption that affected people are essentially passive recipients of assistance, which, as we have seen, often underlies the idea of 'needs assessment', can lead to long-term dependency, and requirements for continued and increased assistance over time (Anderson et al., 2012).

Here is an important reminder that the questions that we ask, and the evidence that we use to answer them, are not neutral. As we have seen, evidence always responds to a proposition, and someone has created that proposition, on the basis of certain requirements and assumptions. Our choice of questions, sources and methods not only reflects the humanitarian system that we have, it also creates the humanitarian system of the future.

IN BRIEF:

- There is no single set of assessment questions that can be used in all cases and, as a result, there has been great variety in the nature, methods and quality of emergency assessments.
- This problem is being addressed increasingly by moves to standardise methodologies (while leaving room for customisation) and to conduct more joint/coordinated assessments.
- Knowledge of the local context and individual judgement are important elements in most assessments.
- Assessments tend to be seen as one-off events, but are more effective where they are part of a longer term process of information collection and analysis.
- A variety of approaches are in use to improve the *accuracy* of assessments.
- Most assessments use purposive sampling, and this can hamper their *representativeness* of the population as a whole.
- A small number of assessments have used probabilistic sampling: however this requires resources that are often unavailable (primarily trained staff and baseline data).
- In general, assessments are still poor at representing the specific needs of marginalised and vulnerable social groups.
- Many assessments are designed to meet the specific information needs of international organisations, and are structured around their conceptual categories, leading to assessments that are inaccurate or not relevant to the real needs of affected people.

Box 3. Evidence and humanitarian standards

The experience of setting minimum standards for humanitarian response tells us a lot about the availability and quality of evidence in the humanitarian sector.

The teams developing humanitarian standards such as Sphere and the Livestock Emergency Guidelines and Standards (LEGS) generally require two types of evidence. The first of these – the 'benchmark evidence' – is used to develop indictors that will allow the results of a programme to be measured against the standard. If, for example, a standard suggests that people should receive ‹sufficient water to meet basic needs, benchmark evidence will be required to suggest how many litres of water per person is enough, in most contexts. This allows agencies using the standard to measure the effect of their programme against a minimum level of adequacy.

The second type of evidence concerns 'what works'. In the Sphere handbook, evidence of 'what works' is used to suggest key actions: 'the suggested activities and inputs' (The Sphere Project, 2011) that will, in most cases, lead agencies to achieve the standards in their programming. In relation to 'benchmark evidence', the Sphere project based some indicators – notably those related to human physiological requirements – on figures derived from empirical research. For example, the indicator of 2,100 kilocalories per day is based on a series of studies brought together by the Committee on International Nutrition (CIN) in 1995 and subsequently endorsed for use in emergencies by the World Health Organization (WFP/UNHCR, 1997). In many cases, however, this type of evidence was not available: both Sphere and LEGS relied extensively on the direct experience of practitioners and on the opinions of experts who reviewed and discussed the information that was available and came to a judgement (Catley, 2013; Damerell, 2013).

Evidence of what works was also hard to come by. At the time the LEGS project started, there were no peer-reviewed publications that addressed the outcomes or impacts of livestock programmes in humanitarian emergencies, and the evaluations and reports that were available from operational agencies had very little evidential value. In general, evaluations focused on outputs (the number of livestock vaccinated) rather than outcomes and impacts (the effects of vaccination on disease morbidity, and the impact on human livelihoods). They could not, therefore, provide evidence on 'what worked' in terms of addressing humanitarian issues. While this might reflect general methodological difficulties with establishing attribution, it also reflected failures in project design: many of the projects that were being evaluated had not clarified the results – in terms of saving lives and protecting livelihoods – that they expected to achieve.

3.4 Evidence from monitoring

Monitoring – the systematic and on-going collection of specified types of information – takes various forms in humanitarian operations. For the purposes of this review, we have identified four broad types of monitoring activity:

- situational monitoring, which is related to needs assessment and speaks to the Type A propositions outlined in Section 2.2, describing the context and showing whether an intervention is – or will be – required
- programme monitoring, which relates to Type B propositions ('what works') collecting information on the implementation and performance of a humanitarian programme

- organisational monitoring, which demonstrates the performance of an organisation across all its programmes, and
- systemic monitoring, which considers the global performance of the international humanitarian system as a whole.

Situational monitoring

Ideally, humanitarian needs assessments – particularly in areas where emergencies are frequent or chronic – would form part of a more comprehensive process of monitoring and understanding humanitarian need. Rather than being 'one off' events, needs assessments would be triggered by, and mark an intensification of, on-going situational monitoring. However, while there is general agreement on the importance of situational monitoring (see, for example, ACAPS, 2012a; ICRC and IFRC, 2008; Garfield, 2013), 'evidence from on-going [monitoring] assessment is sorely lacking – and obvious gap for evidence based interventions' (Darcy et al., 2013).

Situational monitoring systems do exist: Médecins Sans Frontières (MSF) Belgium runs a disease surveillance system in the Democratic Republic of Congo (DRC) that monitors through a network of 'antennae'; while in Ethiopia, there is a joint nutrition-surveillance system involving government, international and national agencies, designed to spot malnutrition 'hotspots'. However, these monitoring systems tend to focus on specific sectors, and are, in any case, few and far between. As noted in the previous section, despite a history of humanitarian engagement stretching back over two decades, routine monitoring of key contextual indicators does not occur in South Sudan (Poole and Primrose, 2010), Somalia (Slim, 2012) or DRC (Darcy et al., 2013).

This lack of situational monitoring appears to be, very largely, a problem of organisational culture and incentives. Continuous, long term investments in monitoring efforts do not fit well into the shorter-term timelines of humanitarian programming, and even where there has been the opportunity for episodic data collection over a period of years, humanitarian agencies have, in many cases, continually changed the method and data collected along the way. Indeed, few attempts have been made to monitor situations for more than a year or two.

"Organisational programme monitoring systems tend to concentrate on the degree to which inputs have been converted successfully into outputs, rather than on the way in which these outputs have influenced levels of need."

In some (albeit relatively few) contexts 'development' programmes have supported information collection and monitoring activities that can trigger humanitarian responses (Mahmood et al., 2010). Investments in longer term activities of this nature are, however, more likely to occur when the focus of international assistance moves from intermittent provision of emergency relief to consistent support to vulnerable populations. This is the case in Ethiopia, where 'what was once a massive humanitarian response machine is slowly being reinvented as an integrated relief, welfare and poverty reduction machine' (Darcy et al., 2013).

Programme monitoring

At the same time, the lack of effective situational monitoring also links to weaknesses in the second type of monitoring – programme monitoring. The issue here relates to the evidential problem of *clarity around context and method*: around what is being asked, why, and for whom. In the same way that agencies' needs assessments tend not to provide impartial measures of total need in a given situation, organisational programme monitoring systems tend to concentrate

on the degree to which inputs have been converted successfully into outputs, rather than on the way in which these outputs have influenced levels of need. One humanitarian organisation calculated that only between 13% and 32% of its monitoring systems were capturing information on the actual effects of programmes on the humanitarian context (Willet, 2013).

It is understandable that donors – who are often the ones asking the questions that monitoring systems are designed to answer – require information on how money is spent: however, this can, in some circumstances, lead to monitoring that largely overlooks whether a humanitarian programme is actually making a difference to humanitarian need (Argren, 2013). The problem is compounded by the fact that most monitoring is done on an agency by agency basis: as a result, it can be hard to identify the cumulative changes caused by humanitarian programming in any one area.

In fact, programme monitoring shares many of the evidential challenges of needs assessment: it can be difficult to establish indicators; to know if the information collected is accurate and representative; and to compare different types of information. But programme monitoring differs in one important way: it is a continuous, rather than a 'one off' process. This leads to additional evidential challenges, but can also improve the quality of the evidence that is generated by monitoring mechanisms.

We have noted that physical access to insecure environments poses a challenge to the collection of much humanitarian information. This challenge is particularly acute for programme monitoring, which requires frequent and regular access to operations. Limited access can make it extremely difficult to be sure of the *accuracy* or *representativeness* of any piece of information, and of data sets built up over time.

Improved telecommunications can help agencies to access information from insecure areas (see Box 4) and agencies are also using innovative 'human' solutions to address this problem. In eastern Myanmar, the International Rescue Committee (IRC) has joined with other agencies to create a system where, whenever one agency has access to an area, it monitors the projects of all the agencies in the area. This approach decreases the potential for bias on the part of the monitor (and so improves *accuracy*); while increasing the size of the sample, allowing for greater triangulation and for more checks to see whether the conditions monitored are representative. In Somalia, UNICEF has taken a rather different approach to a similar problem, using third-party validation to ensure the quality of the information that is being received from monitoring systems. Where basic quality checks suggest that there may be problems with monitoring information, UNICEF sends hired third-party monitors to conduct a follow up investigation. These monitors are 'blind tasked' making it more difficult for them to fabricate information. Again, the approach has the ability to radically improve the *accuracy* of information (Fetouh et al., 2013).

Humanitarian monitoring tends to include affected people less than some other phases of the programme cycle (Alexander et al., 2013), but it may include them more effectively. The longer term and continuous nature of monitoring activities allows for more significant, less tokenistic inclusion: there is time to establish relationships and enter into more of a two-way dialogue than tends to happen

in needs assessments or in evaluation. The recent growth in feedback and complaints mechanisms, which are often an element of programme monitoring, has led to many organisations experimenting with multiple channels of feedback, aiming to establish systems that are robust, context-specific, and that allow marginalised groups to provide information on their experience of aid (Bonino et al., forthcoming; Fenton et al., 2011). As in needs assessment, there is some evidence that these mechanisms improve the evidential quality of monitoring information (Featherstone, 2013; Laybourne and Obrecht, 2013).

Box 4. Technology and evidence

Advances in information technology hold out the possibility to greatly improve the ability of humanitarians to collect and analyse data and to get evidence used.

Information technology means that data can be gathered rapidly after the onset of a crisis, and regularly as the crisis develops. IT can also make evidence more accessible and easy to interpret. However, without systems to adequately process this data, it can become overwhelming. Humanitarians need to build capacity to analyse, understand and utilise evidence gathered. Where resources are scarce, or capacity is limited, low-tech options may be more effective.

As new technologies further develop their massive potential to support evidence-informed humanitarian action, a variety of other considerations will need to be taken into account. While significant amounts of data can be gathered quickly, not all of it amounts to useful evidence. At the same time, the increased amount of data gathered can raise expectations which humanitarians may be unable to meet. And increased dependency on technology can make organisations vulnerable to the technological failures common in post-disaster contexts.

One key concern relates to the evidential quality of 'crowd-sourced' information (that is, information sourced through large numbers of people using mobile telephony or internet based services). Technological advances make it possible to collect data from new and diverse actors. On one hand this makes it easier to triangulate information and paint a thorough picture; on the other hand the accuracy of evidence, and credibility of some sources can be questioned. In addition, there are questions about representativeness. If evidence is gathered via SMS, for example, the source of information is limited to those who have access to mobile phones.

The Ushahidi Haiti Mapping Project provided precise situational awareness that an evaluation found to be highly relevant to early response. However, few humanitarians acted on this information. They were suspicious of the crowd and concerned about evidential quality and representativeness. The data was unsuitable for the rigidity of information requirements of responding organisations. Many have fears that evidence from crowd-sourced platforms will be biased, or manipulated. Others note that all data has inherent bias, and humanitarian decision-making frequently relies on anecdotal 'evidence'. Biased data may be better than no data at all.

Technology has the potential to improve accountability to crisis-affected people by improving information flow and two-way communication, and to donors by providing clear information trails. It can make gathering evidence more participatory, and easier to conduct at scale. However, it also brings data protection concerns. Geo-tagging (recording the location where

information/images were taken) can be useful in analysis, however it has significant security and privacy risks. Similarly, while digital records save time and reduce duplication errors, many are apprehensive about storing large amounts of data in servers or cloud-based systems. Humanitarians should develop guidelines to ensure information is used ethically and securely.

Another area where system-wide action may be useful is in the development of common platforms. Evidence is most useful when it is easily accessible, and while the wide variety of customised data applications make data more readily comprehensible to their users, they are of limited use to people using another system or platform.

Humanitarians should recognise the opportunities new technology presents to improve the quality and use of evidence, and take an active part in ensuring that, as technology is introduced, it addresses issues of data quality, confidentiality, access and interoperability.

Sources: Avila, R. et al (2010); Coyle, D. and Meier, P. (2009); OCHA (2013b); Smith, G. et. al. (2011); World Vision (2013); IFRC (2013)

Organisational monitoring

An increasing number of humanitarian organisations across the sector, in addition to monitoring individual programmes, are attempting to monitor their overall performance as humanitarian actors, by aggregating country- or programme- level information to provide a global picture. The nature of what is monitored differs from one organisation to another: Mercy Corps considers performance against the organisational mission (Willet, 2013); Action Contre La Faim (ACF) against the OECD DAC evaluation criteria (Guerrero et al., 2013); while Catholic Relief Services (CRS) focusses mainly on output measures (Carr, 2013). These differences are reflected in the nature of the data collected, and the degree to which additional data (beyond that provided by programme monitoring and evaluation systems) are included. While Mercy Corps has been very clear that its 'Mission Metrics' system will use only data from existing monitoring systems, Oxfam GB augments programme reports with interviews of programme staff (Walden, 2013).

These initiatives face some common evidential challenges. Given the very different contexts in which an organisation may work, and the very different types of programme that it may undertake, it can be extremely difficult to establish common categories for measurement across programmes (Carr, 2013). Indeed, even the term 'beneficiary' can have different meanings in different contexts. Where information is collected from a sub-set of programmes, there may also be problems with *representativeness*: to what degree do those programmes represent the organisation's work as a whole?

"It can be extremely difficult to establish common categories for measurement across programmes."

Perhaps the most fundamental challenge, however, and as in other areas, is that of *clarity around context and methods*: how is the information intended to be used and by whom? As one expert has suggested: 'How are we doing is not a very good research question' (Willet, 2013). Agencies produce vast amounts of information, and its analysis requires resources that could be used elsewhere. Organisational performance monitoring can be used for strategy development; for advocacy; for donor reporting; and – where information is disaggregated to country level – for performance management. Given the resource implications of these approaches,

it is important to build organisational monitoring systems to answer clear and explicit questions.

Organisation-wide performance monitoring, if done well, has the potential to support improvement in individual organisations. However, increases in the size, scope and complexity of humanitarian responses over the past decade have led to increased interaction and interdependence between humanitarian organisations, as they attempt to address situations that require large-scale, multi-phase, multi-sectoral responses. In this context, monitoring and improving the performance of individual organisations is an important first step in improved humanitarian response. Ideally, however, we would also be able to monitor the performance of the humanitarian aid system as a whole.

System monitoring

A variety of frameworks and systems exist to monitor humanitarian performance. These tend to concentrate on the performance of certain sets of actors within the broader humanitarian system, such as the Good Humanitarian Donorship initiative, which concentrates on donor performance against specific principles (see DARA, 2011); on performance related to specific elements of humanitarian response, such as The Humanitarian Accountability Partnership, which concentrates on accountability to affected people (see HAP, 2010); or on performance related to specific responses, such as the international response to the Indian Ocean Tsunami (Cosgrave, 2007).

Until recently, however, there has been 'hardly any attempt to assess performance of the humanitarian sector as a whole, on either a one-off or regular basis' (Ramalingam et al., 2009: 73). There is no common, agreed framework for monitoring international humanitarian performance, and while (as we have seen) a large number of agency-specific monitoring and reporting mechanisms exist, they are fragmented, and collect a wide variety of different types of information for different reasons.

In 2009, ALNAP, as part of the Network's Humanitarian Performance Project, suggested that a 'balanced, comprehensive and coherent framework for humanitarian performance could be discussed and agreed across agencies' (Ramalingam et al., 2009: 7) and proposed a humanitarian 'balanced scorecard', as the basis of system-wide monitoring. To date, there has been little progress on the idea of a common performance framework – although recent discussions around humanitarian effectiveness, related to the proposed World Humanitarian Summit in 2016 have seen a renewal of interest in the topic (OCHA, 2013). In the meantime, ALNAP has instigated a process of monitoring the performance of the humanitarian system against the OECD DAC criteria[10] with a biennial *State of the Humanitarian System* report (ALNAP, 2010 and 2012). This process aims to describe the size, shape and performance of the formal humanitarian system; identify trends in performance; and – perhaps most importantly – clarify the key issues and gaps that need to be filled by sector-wide improvement efforts.

10 See Beck (2006) for more details on the OECD DAC Evaluaton criteria as applied to humanitarian action.

While these reports have been extremely well received (John Mitchell, personal communication, June 2013; ALNAP Monitoring and Evaluation Log); the process has also thrown up a series of challenges linked to the collection and analysis of evidence (particularly around Type B propositions: 'what works') related to multiple agencies and organisations in multiple contexts.

The first challenge recalls difficulties noted earlier in the sections on early warning and assessment: it relates to the *relevance* of the report's measurement indicators. The OECD DAC criteria used in the report, such as 'appropriateness' and 'coherence' are conceptual categories with a subjective component, which – in most cases – can only be measured by using a variety of different types of data. To make things more difficult, there is no general agreement on which sets of indicators could or should be used to measure each criterion (and, as mentioned, it may not be possible to establish a single set of agreed indicators, given the variety in contexts in which humanitarians work).

As a result, the selection of indicators for the report relies heavily on expert opinion. ALNAP uses a three-stage approach in an attempt to diminish any potential bias in the analysis. The first draft of the report is compiled by expert researchers, with a good knowledge of the system. This draft is critically reviewed by the ALNAP Secretariat, and then reviewed separately by an Independent Advisory Group, composed of individuals with a variety of perspectives from within and beyond the ALNAP Network. The diversity of background represented by the Advisory Group also helps to ensure that those issues that are 'hard to measure' do not get ignored in the final report (Knox Clarke, 2013a).

The second challenge – related to the *accuracy* of the information and analysis – is the familiar one of the sheer lack of data. For several areas that are monitored, there is very little convincing data available. In the area of coverage, for example, 'the system lacks a hard set of numbers on total people in need of humanitarian aid at any given time' (ALNAP, 2012: 42). Where data do exist, they are generated, in general, by actors from within the system itself, and so are potentially prone to bias. This bias can, to a large degree, be addressed by triangulation, but it is important to triangulate with information coming from actors outside the 'formal system', to challenge any 'systemic' or 'cultural' bias that may exist within the system as a whole. In this respect, the inclusion of beneficiary surveys and of interviews with staff from National Disaster Management Authorities (NDMAs) are important elements of the research methodology.

As mentioned above, any measurement of global performance in the humanitarian sector – whether at the level of the single agency or of the system as a whole – will also face difficulties around *representativeness*: to what degree does the sample of information accurately reflect the whole picture? In previous iterations of the *State of the Humanitarian System*, the reviews of evaluations and key informant interviews have been based on purposive samples, while field visits (for 'ground truthing') and beneficiary surveys have been based on convenience samples. In the next phase of the project, visits and surveys will be selected using a purposive approach, in an attempt to improve the degree to which information gained from these sources can more closely represent humanitarian action as a whole (John Mitchell, personal communication, July 2013).

IN BRIEF:

- Monitoring comprises a variety of related but separate activities: situational monitoring; programme monitoring; organisational monitoring and system monitoring.
- The humanitarian system is weak at situational monitoring: there have been few successful attempts to monitor situations over periods of more than a year or two.
- Performance monitoring can focus too much on outputs, and not enough on outcomes and impact, and it suffers from problems with *accuracy*, particularly in situations where access is a challenge.
- Agencies have found a variety of innovative ways to address these problems several organisations in the ALNAP Network are attempting to monitor their global humanitarian performance. Organisational monitoring – particularly where it uses existing information – is difficult, but clarity around what is being monitored, and how the information will be used, is central to success.
- ALNAP monitors the performance of the whole humanitarian system in the biennial *State of the Humanitarian System* report, but this is challenging – in part because much of the necessary information has not been collected, and in part because of difficulties in generalising from a very diverse set of situations.

3.5 Evidence from evaluations and controlled trials

“Evaluations are more likely to provide robust evidence where they use ‘mixed methods’ approaches.”

Evaluations of humanitarian action are faced with similar challenges to needs assessments and monitoring. These challenges will by now be familiar to the reader: evaluations often take place in data-poor, politicised and complex environments, where physical access is limited, populations are mobile, and where there are many different actors, all of whom wish to confirm their view of what has happened. These contexts make it difficult to build an evidence base that is accurate, representative of the experience of the affected population, and not biased by subjective interpretation.

In general, the guidance for the evaluation of humanitarian action suggests that evaluations are more likely to provide robust evidence where they use ‘mixed methods’ approaches. IFRC’s monitoring and evaluation guidelines, for example, suggest that qualitative data allow for only limited *generalisation*, and can be perceived as having low credibility, while quantitative methods can be costly and ‘exclude explanations and human voices about why something has occurred’. As a result, ‘a mixed methods approach is often recommended that can utilise the advantages of both’ (IFRC 2011: 35). Similarly, MSF’s evaluation unit suggests that ‘usually a mix of qualitative and quantitative methods provides the best results’ (MSF 2012: 7), while guidelines from the World Food Programme (WFP) suggest that ‘as qualitative and quantitative data complement each other, both should be used’ (WFP, n.d.: 23).

In the spirit of these guidelines, WFP and the Office of the United Nations High Commissioner for Refugees (UNHCR) recently completed a series of mixed-method impact evaluations on the contribution of food assistance to durable solutions in protracted refugee situations (Steen et al., 2013). These evaluations

used quantitative data from household surveys,[11] based on random samples of the refugee population, and qualitative data from interviews, focus groups and observation. The synthesis report notes that 'the main guarantor of the validity of the findings is the very broad range of sources and methods used to assemble evidence and its triangulation' (Steen et al., 2013: 3). Probabilistic sampling – which improves the *representativeness* of findings – has also been used in evaluations in pastoralist contexts where lists of households were available (Abebe et al., 2008).

Generally, however, humanitarian evaluation 'uses mainly qualitative methods' (Buchanan-Smith and Cosgrave, 2012). An informal review conducted for the paper of evaluations from 2010–2012 in the ALNAP resources library suggests that the use of mixed methods approaches is uncommon, and that most evaluations have relied on broadly qualitative approaches to evidence generation, and particularly on interviews – often with key informants – and personal observation.

Most evaluations use purposive sampling techniques to identify interviewees. They rely on triangulation of sources (and, to a degree, on triangulation of the observations of different evaluators) to establish *accuracy*. The orientation towards qualitative and discursive approaches is particularly marked in evaluations that are primarily for learning (rather than accountability) purposes, as this type of evaluation emphasises the importance of subjective experience and the participation of key stakeholders in the evaluation process as a precondition for learning and change.

While there is sometimes a tendency to see quantitative approaches to evidence generation as 'hard' and qualitative approaches as somehow 'soft', the use of qualitative approaches should not be an excuse for lack of rigour. The six criteria outlined in Section 2.3 for robust evidence are as important for information gathered through interviews as for data from randomised surveys. Unfortunately, many humanitarian evaluations do not use academically recognised qualitative methods, and fail to meet basic quality standards related to *accuracy*, *representativeness*, and *relevance*. One ALNAP member organisation recently reviewed its own evaluations against BOND's NGO evidence principles (Bond, n.d.). Their findings suggest that there is significant room for improvement:

> *According to the BOND tool, the quality of evidence in the…evaluations was found to be low in almost every category identified. This indicated that whilst the quality of the projects being evaluated may be good, the low quality of evidence meant it was difficult to conclude this from the reports. Many conclusions were drawn from a seeming lack of evidence, and many anecdotes were presented as illustrations of more general findings. The reports contained expressions of opinions, but rarely were examples or evidence given to substantiate. There was often a lack of basic quantitative information (e.g. numbers of men/women/elderly/youth met with), and no convincing analytical methods or methodology was presented. (Parker and Sanderson, 2013).*

These findings may not be representative of all humanitarian organisations, but – again on the basis of an informal review of the ALNAP resources library – they also do not seem to be particularly unusual.

11 Measuring food security through the standard indicators of the food consumption score (FCS), household dietary diversity score (HDDS) and Coping strategy Index (CSI).

> **A further area of particular concern in many evaluations is the lack of a beneficiary perspective.**

Recognising the importance of ensuring the evidential quality of evaluations, several organisations have instituted systems to review evaluation quality in recent years: WFP has such a system,[12] and UNOCHA is planning to implement one (Jock Paul, personal communication, April 2013). At the same time, ALNAP is currently piloting a Guide to the Evaluation of Humanitarian Action, which outlines approaches to conducting more robust qualitative approaches (Buchanan-Smith and Cosgrave, 2013).

A further area of particular concern in many evaluations – despite the heavy reliance on interviews as a source of evaluative information – is the lack of a beneficiary perspective. When Beck and Buchanan-Smith conducted meta-evaluations for ALNAP, they found that almost three-quarters of the evaluations reviewed between 2001 and 2004 had failed to consult beneficiaries, or had included only minimal consultation (Beck and Buchanan-Smith, 2008).

Despite some notable exceptions, evaluations still tend to undervalue the experience of affected populations as a source of evidence: the 2012 edition of the *State of the Humanitarian System* (ALNAP, 2012) concludes that recipient consultation is one of the weakest areas of humanitarian performance, while the Humanitarian Accountability Partnership (HAP) 2013 *Humanitarian Accountability Report* notes that beneficiary engagement is 'extremely rare in evaluation, even if the current trend is to push for beneficiary involvement at this stage' (Alexander et al., 2013: 29). The participatory impact approaches discussed below provide a welcome, if atypical, example of how evaluation can be constructed around the experience of disaster-affected people.

Causality and transferability – two particular challenges for evaluations

In addition to the other evidential tests set out in Section 2, evaluators need to address the challenge of *attribution*. 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 (i.e. Type B propositions). As a result, evaluations need to be rigorous in their approach to *attribution*.

Many evaluations of humanitarian action address this challenge of *causality* by relying on a logical framework. In this approach, a project is designed according to a causal chain, which forms a kind of hypothesis. If certain deliverables are produced (say, a certain number of boreholes producing a specified amount and quality of water) and certain assumptions hold true (people obtain their drinking water from this source and not from elsewhere) then the 'logical' assumption is that there will be certain positive outcomes, such as a decrease in water-borne diseases. In this case, if an evaluation can demonstrate that the deliverables were produced and that the outcomes subsequently occurred, and if interviewees create a narrative link

12 The Evaluation Quality Assurance System, or EQAS, which sets out processes with in-built steps for quality assurance and templates for evaluation products. It also includes quality assurance of evaluation reports (inception, full and summary reports) based on standardised checklists…This quality assurance process does not interfere with the views and independence of the evaluation team, but ensures the report provides the necessary evidence in a clear and convincing way and draws its conclusions on that basis.'(FAO and WFP, 2013).

between the deliverables and the outcomes that discounts alternative explanations, then the deliverable is generally held to have caused the outcome.[13]

This approach, while perhaps imperfect, has tended to dominate wherever evaluations focus on the level of outcomes. It is less useful, however, when the evaluation considers the impact of humanitarian intervention, because the causal chain between deliverable and impact[14] tends to be more complex and ambiguous, and may not have been previously articulated.

Measurement of the impact of humanitarian action is challenging for a number of reasons: lack of capacity; high staff turnover; an aversion to publicising failure; as well as technical difficulties in establishing baselines and control groups, and in disentangling the impact of a single intervention from the broader impact of an operation. All of these work against robust impact assessment and, as a result, evaluations of impact are still fairly rare in the humanitarian sector (Proudlock and Ramalingam, 2009). Where they do occur, the meaning of 'impact' often differs from one agency to another (ALNAP, 2012; ACF, Mimeo). Nevertheless, there has been a recent and significant increase in interest in humanitarian impact evaluation, driven at least partly by a desire to establish which approaches offer the best value for money in a constrained financial environment.

Beyond the humanitarian sector, there is a lively debate on the most valid approaches to establishing robust evidence of *attribution*. Proponents of experimental approaches – generally randomised control trials (RCTs) – argue that they represent the 'gold standard' in establishing causality; or, more modestly, 'the worst form of design except all the others that have been tried' (Bickman and Reich, 2009). Critics point to the cost of RCTs – typically $200,000 to $900,000 (World Bank n.d, quoted in Bradt, 2009) – noting that cheaper, non-experimental approaches are used regularly across a range of scientific disciplines to establish causality beyond reasonable doubt (Scriven, 2009). They suggest that the results of RCTs are really effective only where interventions are 'stable and relatively simple and…produce relatively quick and large effects relative to other potential influences' (Piccioto, 2012). Critics also maintain that RCTs are of limited use in policy-making because they are not generalizable to other settings (Schwandt, 2009) and because they seldom explain how an intervention led to specific impacts (Piccioto, 2012).

We can perhaps conclude that RCTs are certainly useful when answering the specific type of 'PICO' clinical question for which they are designed (Bradt 2009).[15] Beyond the clinical sphere, they will tend to be most effective in establishing *attribution* 'where the causal chain between the agent and the outcome is fairly short and simple and where results may be safely extrapolated to other settings' (Victora et al., 2004, quoted in Dijkzeul et al., 2013).

13 The approach has interesting echoes of the rationalist, rather than empirical, understanding of truth, relying as it does on a logical sequence, rather than on direct observation of the causal link.

14 Often understood as the final link in a causal chain: 'Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended' (OECD-DAC, 2002).

15 PICO stands for Patient (or population) receiving intervention; Intervention under consideration; Comparison (the alternative intervention being considered); clinical Outcomes being sought. Consideration of these four factors leads to the creation of a 'testable' question. PICO questions are often of particular relevance for issues such as determining which therapy will be most effective.

“We can expect the number of controlled trials, systematic reviews, and other approaches that prioritise experimental methodologies to increase although there will be many situations where experimental approaches are not desirable or feasible.”

There have been some attempts to use RCTs in the humanitarian arena to provide evidence of impact. DFID has funded a RCT in Malawi to test different compositions of ready-to-eat food in the treatment of severe acute malnutrition (Kerak et al., 2009 in Buchanan-Smith and Cosgrave, 2012) and ACF has conducted RCTs on the impact of adding ready-to-use food to supplementary feeding programmes on child malnutrition (Huybregts et al., 2012). In Liberia, the IRC has conducted an impact evaluation of a community-driven reconstruction programme using an experimental design (Fearon et al., 2008): an interesting – and to date fairly unusual – example of using experimental approaches in humanitarian work outside the health and nutrition sectors.

We can expect the number of controlled trials, systematic reviews, and other approaches that prioritise experimental methodologies to increase, supported by organisations such as 3IE and EvidenceAid, who are attempting to increase the rigour and sophistication of evidence generation in the humanitarian sector. As this work progresses, there will be much to learn from early attempts. ACF experience in Chad confirms that RCTs in humanitarian contexts require a long period for design and implementation, and are vulnerable to changes in the situation of the groups taking part. Their success depends on highly skilled researchers (in this case, through a partnership with the University of Ghent and the institute of tropical Medicine, Antwerp) and on frequent and effective communication with participants, the population at large, and the organisation’s staff on the ground (Puett and Salpeteur, 2013). Given the high costs involved it is important to select topics that are highly strategic: where, in effect, the cost of expenditure is balanced by the cost of *not* doing the research (Chloe Puett, personal communication, April 2013).

At the same time, there will be many situations where experimental approaches are not desirable or feasible. Both Stern, in a recent paper for DFID (Stern, 2012) and Rogers, in a paper for Interaction (Rogers, 2012) suggest alternative designs and methods, including ‘quasi-experimental’ approaches, case-based approaches, and theory-based approaches. Both point to the desirability of using a variety of approaches to consider causation and note the importance, in the humanitarian context, of considering the degree to which interventions contribute to changes, rather than attempting to attribute change solely to any single intervention.

Increasingly, humanitarian practitioners are experimenting with a range of methods to address the issue of *attribution* (Box 5). The WFP/UNHCR impact evaluations, mentioned above, used ‘opportunistic’ comparison groups – such as official and unofficial groups of refugees, or camps that received different quantities of assistance. In the absence of a formal logical framework for the programme, the evaluators also reconstructed the logic of the interventions, to see whether the input (food aid) really did lead to impact in the way originally envisaged.[16] The ECB project has developed and tested a methodology that considers the contribution of humanitarian interventions to change, using descriptive statistics combined with interview data (Few et al., 2013).

As noted above, impact assessment is one of the few areas where there have been concerted attempts to build evaluation around the perceptions of affected people.

16 Interestingly, the evaluation was able to conclude that ‘major assumptions concerning refugees’ use of food assistance have not held’ (Steen et al., 2013: 14).

The Feinstein Institute of Tufts University has produced guidance on participatory impact assessment: this notes the practical and ethical difficulties of establishing control groups to test *attribution* and instead focuses on the use of participatory tools to assess the relative contribution of project and non-project factors to change (Catley et al., 2008).

The participatory approach is not, however, just a way to get around practical difficulties. Participatory evaluations recognise that impact – being largely about the longer term and often unexpected consequences of an intervention – do not fit neatly into externally-generated project logic, and are, in general, best understood through the experience of aid recipients themselves.

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 (see, for example, O'Hagen and McCarthy, 2012). A similar approach was used by FAO in Somalia: here 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).

Box 5. Research designs for investigating attribution

Experimental design:
In an experimental design, participants are assigned randomly, either to a group that receives programme services or to a control group that does not, and that serves as a 'counterfactual'. Outcomes from these two groups are then compared. The design of the experiment allows any difference in outcome between the recipient and control groups to be attributed to the services received.

Quasi-experimental design:
Quasi-experimental studies also aim to demonstrate attribution by comparing outcomes, but they do not involve the random assignment of participants to groups. Instead, they compare outcomes for groups who received services and for similar groups who did not (a 'natural experiment'); or for one group before and after an intervention.

Theory-based approaches:
These approaches do not attempt to demonstrate attribution by comparing recipient and counterfactual groups. Instead, they test the underlying theory of causation by which programme designers expect certain activities to lead to certain results. In a theory-based approach, the series of assumptions in the programme design which link input, context and result are treated as hypotheses, which can be tested using a variety of methods, quantitative and qualitative.

Case-based approaches:
Case-based approaches rely on a study of what actually happened in specific cases: identifying the factors that led to certain outcomes, and then comparing them within cases, or between cases, in order to make 'analytical generalisations'.

Sources: definitions based on Stern et al. (2012); Morra Imas and Rist (2009); Leeuw (2012).

IN BRIEF:

- While much of the guidance for humanitarian evaluation points to the benefits of using mixed methods approaches, most evaluations are conducted using only qualitative methods.
- Typically, humanitarian evaluations use purposive sampling to try to ensure their *representativeness*, and triangulation to increase *accuracy.*
- However, many humanitarian evaluations fail to make good use of these qualitative methods and, as a result, the quality of the evidence that they present is often poor.
- Many evaluations fail to include the perceptions of people affected by crises in any meaningful way.
- Some of the most effective work on structuring evaluation around the perceptions of affected people has been seen in the area of impact assessment.
- *Attribution*/contribution are particular challenges for evaluations, which aim to investigate causal links between a programme and a subsequent situation.
- There has been some recent growth in 'experimental' approaches to establishing *attribution*, and agencies are also beginning to consider a variety of other, non-experimental approaches to the assessment of contribution.

4

Is evidence currently used to guide humanitarian operations and policy-making?

4.1 Quality does not guarantee use

The previous section considered the quality of evidence generated within the international humanitarian system. If decisions are to be informed by evidence, it is, of course, vital that evidence is of the highest possible quality. But quality alone does not guarantee that the evidence will be used. Decision-making processes are rarely, if ever, entirely rational: they are influenced by a variety of organisational and political considerations, and by the biases of individual decision-makers.

If evidence is to be of use in informing humanitarian action, then we need to address not only the process of information collection and analysis, but also the process of decision-making. The experience of the medical profession is instructive here: the move towards 'evidence-based medicine' that took place in the 1990s owed much to changes in organisational procedures and changing attitudes among individual practitioners (Anthony Redmond, personal communication, October 2012): 'evidence-based decision-making often requires no additional scientific data per se, but rather…financial resources and political will' (Bradt, 2009: 3).

With this in mind, this section considers the degree to which different types of evidence and knowledge have been used by humanitarian decision-makers, and investigates the factors that appear to constrain, and to support, the use of evidence in decision-making.

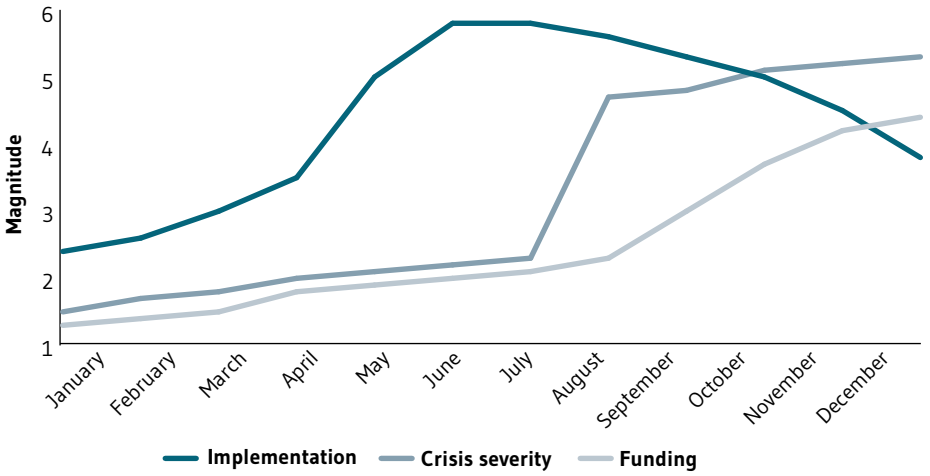
4.2 Use of early warning evidence

"The problem of a disconnect between early warning information and actual responses remains at the heart of policy and decision-making."

As noted in Section 3, considerable progress has been made over the past two decades in the generation of timely and accurate early warning evidence, making it possible to formulate Type A propositions with some confidence in many contexts. The use made of such information, however, has been more problematic, particularly in relation to famine and food insecurity. The problem of a disconnect between early warning information and actual responses has long been recognised (see Buchanan-Smith and Davies, 1995) yet the problem remains at the heart of policy and decision-making in, for example, slow-onset food crises (Levine et al., 2011).

Some of these issues could be seen in relation to the 2011 famine in south central Somalia. Figure 1, taken from the IASC Real Time Evaluation (RTE) of the Somalia crisis response, sets out the relationship between crisis severity, response funding and response implementation.

Figure 1. Schematic representation of response to the Somalia famine of 2011



Source: IASC Real Time Evaluation of the international response to the Somalia Crisis 2011 (Valid International, 2012).

The most striking feature of Figure 1 are the gaps between, first, crisis severity[17] (the green line) and funding availability (the light green line) and, second, between funding and response implementation (the dark green line). Strong early warning evidence of the impending crisis was available from the last quarter of 2010 onwards (from FEWS Net and FAO's IPC), but a major increase in funding only came with the declaration of famine in July 2011. This was a primary factor in what the authors of the real time evaluation refer to as 'a systemic failure of early response' which had two aspects:

- the failure of prevention action to tackle the proximate causes of vulnerability through urgent livelihoods intervention that would build short-term resilience and reduce the need for relief
- the failure of scaled-up early relief to tackle the most acute symptoms of the crisis at the moment such assistance was most badly needed in early to mid-2011 (Valid International, 2012: 37).

In the event, the planning for a massively scaled-up response only took place in July 2011, after the most acute phase. The authors of the RTE noted: 'Given the lead times involved, this needed to have happened in January/February at the latest if the preventive and relief agendas were to be addressed at the time required.' By the time programme responses were being scaled up, the crisis was already well past its peak.

The situation in South Central Somalia in 2010–2011 was complicated by political and security considerations. Large parts of the affected area were, at the time, controlled by the Al Shabaab group. Major donors refused to allow support to go to Al Shabaab, while Al Shabaab itself prevented many agencies from operating in the region. However, the pattern of early warning and late response had occurred before, where these constraints had not been in place to the same degree (Hillier and Dempsey, 2012) and was repeated in other parts of the Horn of Africa. The peculiar circumstances of Somalia, in other words, do not explain the phenomenon.

17 'Crisis severity' as represented here is based on a basket of indicators including levels of acute malnutrition and market food prices. These three variables are recognised as being both difficult to quantify and incommensurable, so that Figure 1 is indicative only.

Nor was the problem primarily one of the availability or quality of evidence. As noted above, early warning evidence was timely and seen, in general, as being of good quality – although the validity of some of the data and analysis was open to question (Levine et al., 2011). 'Ultimately, the early warning systems performed, but decision-makers chose not to respond' (Hillier and Dempsey, 2012: 14).

There seem to be a variety of reasons for the failure of the international humanitarian system to make use of the early warning evidence. These constraints relate to the nature of the evidence available; the decision-making processes prevalent in the international humanitarian system; and the broader operational processes of the system as a whole.

The first constraint – the nature of the evidence available – relates to the fact that, even when organisations knew that the situation was likely to lead to significant and negative impacts on the lives and livelihoods of people in the region, they did not know what to do with this information: how to turn that knowledge into effective responses (Levine et al., 2011). While there are documented examples of effective preventative action in situations such as these (see the discussion on early intervention in drought-related emergencies in Section 4.4), most organisations do not appear to have been aware of these examples, or to have experience in using them, having focused, historically, on responding after a crisis is already evident. As a result, many humanitarian organisations find it difficult to turn early warning information into concrete programmes.

Given the humanitarian propositions outlined earlier, this is an important finding: in the Horn of Africa, it was not possible to use evidence for a Type A proposition (that a problem exists) without also having evidence for related Type B propositions (that a certain approach is effective in addressing the problem). Early warning may have been more effective had it been accompanied by examples of programmatic 'good practice' from those agencies with experience in this area.

This was not, however, the only reason for inaction. The failure to respond in Somalia also highlights the second constraint: weaknesses in the way decisions are made in the humanitarian system. Hillier and Dempsey note that 'many people ''on the ground'' in the region…were aware of the impending crisis and trying to set alarm bells ringing in January and February 2011, [but] they were not always able to get traction ''further up the chain'''(Hillier and Dempsey, 2012: 4). The global and atomised nature of the humanitarian system means that decisions often require consent from a large number of actors in different places and in different organisations; and that it is not always clear where and by whom decisions should be made (Knox Clarke, 2013b). Moreover, individuals in different positions are often working under different sets of incentives, and liable to give different weight to the same evidence.

In the case of Somalia, it appears that (to put it simply) field staff were motivated to heed the early warning data by the deteriorating situation they saw on the ground. Members of the HCT in Somalia, on the other hand, were aware of the difficulties of raising funds, and were more reluctant to engage with the information (Slim, 2012).

Some decision-makers in Ethiopia considered that the needs identified by early warning systems were 'too small' to catalyse an effective donor response (Majid, 2011).

Meanwhile, donors in western capitals were mindful of the need to make the best use of taxpayer resources, and were uncomfortable releasing funds until there was evidence – in the form of nutritional data – that a humanitarian crisis was *already* occurring (Hillier and Dempsey, 2012; Levine et al., 2011). At the same time, governments in the region were unwilling to call for international assistance unless it was absolutely necessary, as to do so would be seen as an admission of failure on their part.

Across the board, the actors had to decide how much attention to pay to specific pieces of evidence in a situation of uncertainty, and these decisions were strongly influenced by other information and organisational incentives. This is a common problem in humanitarian operations – many authors have commented on high levels of risk avoidance in the humanitarian system (see Buchanan-Smith and Scriven, 2011), which, it could be argued, tends to shift the risk away from humanitarian organisations and onto the shoulders of the disaster-affected people themselves.

“High levels of risk avoidance in the humanitarian system, tends to shift the risk away from humanitarian organisations and onto the shoulders of the disaster-affected people themselves.”

In 2010–2011, these decision-making problems were compounded by a lack of ‘triggers’ (Levine et al., 2011): clear statements of levels of need – particularly in sectors such as water and sanitation – that would automatically force a decision on response. In effect, early warning systems were part of a larger decision-making process: unfortunately, they appear to have been significantly better developed than other elements of that process. There was a particular lack of clarity over key issues, such as who should make decisions; the relative importance of different criteria in evaluating evidence to make decisions; and even over when and whether decisions should be made at all.

The third set of constraints to the use of early warning evidence concerned the broader operational processes that convert funding into action. In general, the current release of funds and initiation of activities is tied to needs assessment, rather than to early warning mechanisms (Levine et al., 2011). Needs assessments – concentrating on actual and immediate needs rather than on anticipated needs – tend to take place some time after early warning assessments, when the situation has already deteriorated. There is then, typically, a delay of several months before operations begin on the ground. From this perspective, early warning evidence is not used to make decisions about intervention because it is not formally designed to lead to such decisions.

To remedy this, Levine and colleagues propose a series of concrete preparedness actions, in areas such as recruitment and procurement, that agencies could take on the basis of early warning evidence to shorten their response time. Similarly, Hillier and Dempsey recommend a series of ‘No regrets’ actions that should be triggered by early warning, including capacity building and prepositioning of stocks.

IN BRIEF:

- Despite improvements in the quality of early warning systems, evidence that a crisis is likely to occur often fails to translate into action, as demonstrated in the failure to respond to such evidence in the Horn of Africa, and in particular in South Central Somalia, in 2010–2011.

- The failure to use early warning evidence has a variety of causes, but one key factor is that the generation of early warning evidence is not part of a larger, structured process of early warning and response:
  - the current 'business process' of humanitarian response links response to assessment activities, not to early warning, and the role of early warning evidence in generating a response is not yet clear
  - it is rarely clear who is meant to make decisions related to early response, and so nobody has clear responsibility for these decisions
  - it is rarely clear what actions should be taken on the basis of early warning: humanitarian organisations have more experience in responding to crises than to anticipating them.
- In addition to these structural and process failures, organisational incentives (and in particular, a desire to avoid the risk of 'wasting money') discourage any action that takes place before a crisis exists.

4.3 Use of assessment and monitoring evidence

While the place of early warning in the decision-making process is often unclear, no such confusion exists over the role of needs assessments, which are designed specifically to provide evidence for use in decision-making. Typically, this evidence is concerned with Type A propositions related to the existence of needs, and allows decisions to be made on whether assistance is required; how much;[18] for whom; and for how long. Assessment is often considered by at least two sets of decision-makers: agency staff, who use it for initial programme design and funding requests; and donors, who use it to determine whether funding requests are legitimate. In addition, humanitarian practitioners often expect to be able to use information from needs assessments as evidence to support decisions around Type B propositions: the best type of assistance to provide in that specific context. This latter use is, however, problematic, and we will return to it in a moment.

A review of the literature suggests that, when it comes to decisions related to Type A propositions (whether to intervene, where, and at what scale) the use of needs assessments is inconsistent, and that where assessments are used, the evidence that they provide is often only one of several factors that influence decisions. This might at first appear surprising – assessments are, after all, designed primarily as decision-making tools – but is explained by a variety of constraints related both to the assessments themselves and to the nature of humanitarian decisions.

As we have seen, the evidential quality of humanitarian needs assessments is uneven, and – as assessments are not often conducted by impartial actors, but by the same agencies who will then use them to justify funding requests – they are often treated with some scepticism by donors (Bradt, 2009; Darcy et al., 2013). From this perspective, addressing the problems with evidential quality outlined in Section 3.1 (above) would, undoubtedly, improve the use of assessments in decision-making.

A second constraint inherent to the assessment process itself (rather than to factors largely beyond the control of the assessor) is that of timeliness. For assessments to

18 Although, as we have seen, a common failing of assessments is to assume that assistance is required, and to focus more on how much assistance should be provided.

influence decision-making, the information has to be accessible when the decision is being made.

Reviewing the uptake of a rapid initial assessment of needs by ACAPS following the Haiti earthquake of 2010, Darcy and Garfield note the effect of delays in making the results of the assessment available to clusters and other decision-makers. They conclude that 'it is not clear whether, had the analysis been made available sooner, it would have informed decision-making around the revised Flash Appeal or Cluster plans. What is certain is that, even assuming the analysis was relevant and credible, it arrived too late to inform initial planning decisions' (Darcy and Garfield, 2011).

A phased approach to assessment, such as that outlined in the MIRA guidance (IASC, 2012a) can help to address this constraint, by ensuring that some information is available within 72 hours, with increasingly sophisticated analyses provided in the following weeks and months. Obviously, the information that can be produced in the first 72 hours is fairly limited: needs assessments of rapid-onset crises, perhaps more than any other area of humanitarian endeavour, challenge practitioners to find a balance between evidential quality and speed – and the objective must be to find evidence that is 'good enough' to support decision-making (Lars Peter Nissen, personal communication, December 2012).

Time is at a premium in many humanitarian contexts. This not only stops decision-makers waiting for evidence, but it also deters them from wading through long assessment reports. If evidence is to be used, presentation matters. ACAPS have found that short reports that present information graphically are much more likely to influence decision-making (ACAPS, 2012). A paper by the Humanitarian Policy Group (HPG) on Humanitarian Diagnostics notes that, in general, 'the way in which information is presented can be crucial to its uptake and use by decision-makers', and recommends that assessments be short, succinct, and include 'killer facts' to catch the attention (Darcy, 2009: 11).

In some cases, particularly when the decision-maker is within the organisation that conducted the assessment, or is generally pre-disposed to using assessment information, ensuring that reports are robust, clear and timely may be enough to ensure that the evidence within them is used. In most cases, however, the degree to which assessment is used will also be determined by factors that relate more to the decision-maker – and the context within which they make decisions – than to the assessment itself. In particular, assessment evidence is more likely to be used when it suggests a course of action that is possible or desirable for decision-makers, and when it fits their (often unconscious) assumptions about the situation.

Recently, several authors have noted the importance of 'path dependence' in determining the degree to which assessment information is used in decision-making (Darcy et al., 2013; Maxwell et al., 2013). Essentially, the idea of path dependence recognises that humanitarian decisions are not made in a vacuum, but in a context where certain elements are already fixed (or at the very least, hard to change). Most decisions are influenced strongly by decisions that have already been made about organisational strategy, capacity, or allocation of funds.

When deciding on issues such as the size, location and nature of interventions, 'most decisions [taken within agencies] appear to be made within quite tight parameters:

the range of options being limited by previously decided questions about strategic priorities, available resources, and so on' (Darcy et al., 2013: 7). Similarly, when deciding on what type of intervention to perform, agencies do not start with a blank sheet of paper. Rather, 'response choices are also driven by the capacity and organisational ethos of the implementing agency' (Maxwell et al., 2013: 7).

Even where there is more latitude for decision-making – where prior decisions do not determine resource allocation – a needs assessment is 'only the beginning of the range of information, influences and considerations that ...[decision-makers] must weigh up' (Poole and Primrose, 2010: 5). Poole and Primrose note that donors, in allocating funding, need to take a variety of factors into account, and that these may have a legitimate influence on decisions. Even if assessments suggest a need to intervene in a certain place and at a certain scale, donors may choose not to do so because needs are greater elsewhere, or because poor security conditions or a lack of trustworthy partners make it unlikely that an intervention would succeed. Political and media considerations may also play a part in the decision.

The reality of humanitarian decision-making, then, is that it requires a conscious weighing up of different sets of evidence – in this case evidence related to need, to capacities and to security conditions – in order to come to a judgement. In these circumstances, good decisions are perhaps better understood as being 'evidence-informed' than 'evidence-based': as different pieces of evidence will tend to support competing – but equally true – propositions. At the same time, however, there are often a series of less rational, and often less conscious, elements that influence decision-making.

“Good decisions are perhaps better understood as being 'evidence-informed' than 'evidence-based': as different pieces of evidence will tend to support competing – but equally true – propositions.”

Working under time pressure, in situations where there are large amounts of information to consider, decision-makers will tend to resort to 'short-cuts' – simple assumptions, generalisations and rules of thumb, which make it easier to order and make sense of information: psychologists refer to these as heuristics. These heuristics serve as a filter, and have the general effect of decreasing the degree to which people will notice and engage with unexpected, unusual or challenging information (Fiske and Taylor, 1991; Macrae et al., 1993), particularly when decisions are required and time is limited (Bohner et al., 1995). The general human tendency to following existing 'mental models' – even when they are challenged by new evidence – has been noted in the humanitarian sphere (Darcy, 2009; Maxwell et al., 2013; Knox Clarke, 2013b), and appears to influence the uptake of assessment evidence in three specific ways.

The first relates to what Peter Walker has called 'starting models'. We have noted already that needs assessments are often built on a series of assumptions – the most important being the often incorrect assumption that there is no local capacity to respond to the crisis. The same can be true of humanitarian operations more generally: they begin with assumptions – or starting models – about a situation, and this model becomes a type of heuristic: a way to filter and order information. In this context, evidence from needs assessments that disproves the starting model will tend to be ignored. An example of this is the situation in Afghanistan, where, after the military intervention in 2001, 'the model was the conflict is over, it's OK to work with the government – they are not a combatant' (Walker, 2013).

Starting models are so durable because they tend to be shared by a large number of people in the humanitarian sector. This brings us to a second heuristic: that of social knowledge. Where a narrative has become widely accepted, information that disproves it will struggle to gain attention. Poole and Primrose point to 'the widely held conceptual narrative of a rapid transition to development' (Poole and Primrose, 2010: 1) in South Sudan, which made it difficult to highlight humanitarian needs. A similar dynamic can be seen in the Democratic Republic of Congo (DRC), where a 'post-conflict narrative' has led to a situation where 'the DRC Humanitarian Action Plan is now routinely underfunded, yet humanitarian needs remain as high as ever' (Darcy et al., 2013: 10).

Finally, and perhaps most pervasively, we have the 'devil you know' heuristic, where decision-makers will tend to interpret situations through previous experiences. These experiences, in themselves, are a type of evidence – but the evidence is not always relevant to the new situation, and not always robust. Decision-makers will also tend to default to 'tried and tested' ways to respond to circumstances (Maxwell et al., 2013), and – where they do not – will demand far higher levels of evidence before supporting 'new' courses of action (Darcy, 2009). This tendency to discount information that suggests unfamiliar conditions or courses of action can, of course, be supported by organisational incentives to avoid risk-taking.

All of the factors outlined above – path dependence, competing priorities and various heuristics – apply to decisions related to both Type A propositions (whether a response is required) and to Type B propositions (what type of response to conduct). However, there is also a constraint to the use of assessments that is specific to Type B propositions, and that is the focus of Maxwell's work on response analysis: a disconnect between 'the kind of information typically provided by assessments and the kind of information that response choice requires.' (Maxwell et al., 2013: 1).

The problem here is not the quality of evidence, nor is it that evidence cannot be acted on, or is ignored. It is rather that the evidence required to act is incomplete. The process of needs assessment is not collecting all of the evidence that decision-makers need, in much the same way as early warning systems are not able to provide information on what humanitarian organisations should do to prevent a crisis. To a degree, this might be a reflection of the 'silos' that exist between different information specialists in the humanitarian system: needs assessment people concentrate narrowly on needs, and leave 'response' to programme staff, in the same way that evaluators often work separately from the people who manage programmes (Argren, 2013; Hallam and Bonino, 2013).

Whatever the cause, the consequence is that decision-makers do not get all of the information that they need. The importance of collecting evidence, not just for general propositions, but to address specific questions that are of concern to decision-makers, is a theme that we will continue to explore in the next section.

IN BRIEF:

- The use of needs assessments in decision-making is inconsistent.
- This is, in part, a consequence of problems with the assessments themselves:

they are not always seen as being impartial; they are not always produced in time; they can be inaccessible and 'hard to read'; and they do not always include the information that agency staff need to plan programmes.

- It is also a consequence of the way in which decisions are made:
  - decision-makers often have limited latitude in decision-making – many choices have already been ruled out by organisational strategies or mandates
  - they also need to take a variety of information into account when making decisions – information on needs is only part of a larger set of information
  - decision-makers are influenced by unconscious biases, which may lead them to ignore information from assessments.

4.4 Use of evaluation evidence

The number of humanitarian evaluations has grown significantly over the past decade. The ALNAP resources library – which is by no means comprehensive – contains more than 1,200 evaluations of humanitarian action. As such, it is one of the largest single sources for evidence on 'what works' (and what doesn't) in international humanitarian response.

As Telford and Cosgrave note in the Tsunami Evaluation Coalition (TEC) synthesis report, in a context where many agency reports 'concentrate on successes and ignore or gloss over failure… [and] media coverage tends to concentrate on single dramatic instances rather than a balanced review of overall quality, [t]he most detailed information on agency performance may be obtained from agency evaluation reports' (Telford and Cosgrave: Synthesis 2006: 108).

Although the number of evaluations has been growing, there appears to be some scepticism as to the degree to which these evaluations are actually used. The use of evaluation evidence relates both to operational practice and to policy-making in the humanitarian sector. Sandison (2007), following Patton (1997), describes three primary uses of evaluation findings:

i. judging the merit or worth of a programme (e.g. accountability to stakeholders; to inform funding decisions)
i. improving a programme (e.g. on-going learning and development)
ii. generating knowledge (which can be used in subsequent programmes and elsewhere).

As Sandison notes, both uses (i) and (ii) can be seen as 'instrumental': the evaluation is used to make direct changes (often based on recommendations from the evaluators) to the programme, sector or systems being evaluated. Most evaluations in the humanitarian sector are designed primarily for this type of instrumental use.

The third type of use (iii) is more general, and less direct. Here the information contained in the evaluation adds to a general store of knowledge on a topic, and is used to make changes and improvements indirectly, and at a distance. While this area may be seen as more properly the domain of research, evaluations often make a significant contribution to generating knowledge (see for example ALNAP Lesson Papers: (Hedlund and Knox Clarke, 2011; Sanderson and Knox Clarke, 2012).

In the terms of this paper, the first two kinds of use are likely to have a lower evidential threshold than the third. While the evidence to inform changes in a specific programme or practice should (ideally) meet minimum standards of *accuracy*, *representativeness* and *relevance*, and address issues of *attribution* effectively, it does not need to be *generalisable.* The third category is different, and the extent to which it is taken as contributing to knowledge beyond the specific situation evaluated relates to the degree with which all of the criteria for evidential quality – including that of *generalisability* – are met.

The instrumental use of evaluations

With regard to the instrumental use of evaluations, Sandison's conclusions are fairly negative: 'Only a minority of evaluations are effective at introducing evident changes or improvements in performance' (Sandison, 2006: 91). She adds, 'instrumental use is the least likely form of utilisation' (ibid.: 121). However, she also notes that, 'we do not know even how many evaluations are conducted, let alone how many are used' and so 'the source of concern regarding non-use in the sector is mostly anecdotal' (ibid.: 91).

In fact, while it is not hard to find examples of evaluation recommendations that have been ignored, and while 'in general the literature describes an inconsistent and, in some cases, a dismal record of evaluation use' (Sandison, 2006), the picture is by no means wholly negative. Professional evaluators contacted during the preparation of this paper pointed consistently to recommendations that had been implemented. The report of the TEC, for example, is claimed to have led to improvements in surge capacity across the system and to have provided impetus to the work of the Needs Assessment Task Force. Similarly, the Second Cluster Evaluation led to an increased focus on local authority engagement in international responses. And evaluators who participated in ALNAP's Evaluation Community of Practice, and contributed to the ALNAP Study *Using Evaluation for a Change* (Hallam and Bonino, 2013) were able to point to numerous situations where evaluations were used, and identified several conditions that appeared to support evaluation use.

The picture is similar when one looks at the (relatively few) quantitative records of the implementation of evaluation recommendations. The management response matrix to OCHA's intermediate review of the Central Emergency Response Fund showed that, in the year after the review, 50% of recommendations were implemented (OCHA, 2007). When WFP studied the degree to which evaluation recommendations had been taken up, they found that 54% had been implemented and 65% had been included in successor documents. WFP found that recommendations were more likely to be implemented where they were operational, rather than strategic, and where their implementation required action from only a limited number of people. Broader recommendations, or those that required coordination with partners or headquarters units, were less likely to be implemented, as were recommendations with intangible benefits or those that implied criticism of WFP staff (WFP, 2005).

“The evidence provided by humanitarian evaluations *is* often used to make 'instrumental' changes to funding or to programmes, but in a highly selective manner.”

These observations suggest that the evidence provided by humanitarian evaluations *is* often used to make 'instrumental' changes to funding or to programmes, but in a highly selective manner. In determining whether an evaluation is used, the quality of

the evidence may matter less than the degree to which any given recommendation is easy to implement.

Over the past decade, many humanitarian organisations – including DFID, SIDA, UNICEF and WFP – have attempted to identify ways in which they can improve the uptake and use of evaluative evidence. ALNAP has also published four papers on the topic – Hallam and Bonino (2013); Hallam (2011); Sandison (2005); and van de Putte (2001) – based on these experiences and those of other Network members. This research suggests that evaluations are more likely to lead to changes in programme implementation or funding where there is already interest in, or discussion around, the performance of a programme; where the production of the evaluation coincides with a 'window of decision-making' (such as a programme extension); where results are communicated in an appropriate and accessible format to decision-makers; and where mechanisms exist for 'follow-up'.

Perhaps the single most important lesson to emerge from these studies, however, is the importance of engaging operational decision-makers in every step of the evaluation process: from selection of the evaluation questions, through information collection, to implementation and follow-up. This helps to ensure the *relevance* of the evaluation to operational needs, and builds ownership of findings. At the same time, the close involvement of programme staff raises questions around the objectivity of evaluation findings and – where objectivity is seen as an important element of methodological rigour – can lead to concerns around the evidential quality of the evaluation. As one author notes, 'there is generally a tension between the independence of evaluation departments and their success in engaging users of evaluation' (Foresti, 2007).

Recent developments in humanitarian evaluation have tended to incorporate some, or all, of these approaches in an attempt to increase the likelihood that evaluations will be used. There has been a growing interest in Real Time Evaluation (RTE) in an attempt to produce information on the progress of an operation that can be used for immediate 'course correction'. In April 2011, the IASC included Inter-agency Real Time Evaluations as a necessary component of all system-wide (level 3) emergencies. Here, not only was information to be made available in a timely manner, but the evaluation exercise was tied explicitly to decision-making, as the RTE was designed to 'inform the Principals' meeting at the end of the 3-month activation period' (IASC, 2011). The IASC is now moving towards implementing Real Time Operational Reviews, which will be implemented primarily by Humanitarian Country Teams in the first instance, a move that may be intended to increase country ownership of the results.

In Haiti, Groupe URD have implemented what Grünewald calls 'Real Time Evaluation plus'. Here, a team conducts a series of evaluations of the same project over a period of two years, working closely with the project team throughout the process. The later evaluations concentrate largely on identifying progress made on the recommendations of the previous missions and the identification of new challenges. As Grünewald explains: 'This leads to a powerful dialogue between the evaluator and the programme staff that goes on over the life of the project… the evaluator loses a degree of their independence (although hopefully not their objectivity) in order to become an agent of change… the gains in improvement – which is, after all, the main purpose – make this worthwhile' (Grünewald, 2012: para. 3).

ACF is also encouraging dialogue between evaluators and field staff, in an attempt to increase the use of lessons from evaluations. The organisation has changed its evaluation process to ensure that evaluators routinely identify best practices as part of their work. Programme staff are asked to consider, discuss and elaborate on these best practices, which are then included in a learning review, and disseminated across the organisation (see the ACF Learning Review, 2011). As a result, evaluative objectivity is maintained and the crucial link between evaluation and organisational learning is strengthened significantly (Guerrero, 2012; Allen, 2012).

General evidence and policy – the use of evaluations and research

We turn now to look at non-instrumental uses of evaluations – and of research more broadly. The overarching question here is how – and to what degree – does evidence from humanitarian action in specific contexts inform humanitarian practice at other times and in other places. To what degree, for example, can good quality evidence from Afghanistan in 2000 influence programming in Zimbabwe in 2012?

There are a variety of channels through which evidence can inform practice. For example, syntheses of evidence, such as the 'lessons papers' mentioned above, might be read by individuals in the field, and inform their actions directly. Evidence from evaluations may be used to create standards, which then guide action in programme design or delivery. In some cases, evidence may be used to develop decision-support mechanisms or software and will, therefore, have a direct influence on decisions made using this software.

Perhaps the most important route from evidence to practice, however, is through policy: evidence informs policy, which in turn guides practice across an organisation. In this section we consider the ways in which evaluative and other evidence has influenced policy development, concentrating particularly on four examples: cash-based programming; the introduction of minimum standards; the consideration of cultural factors in humanitarian programming; and early intervention in drought-related emergencies.

Cash-based programming

One of the most significant policy developments in the last decade has been the increased acceptance and support of the use of cash in place of distributions of food and other goods. The humanitarian assistance policy of ECHO (the humanitarian aid and civil protection department of the European Commission), along with the organisation's guidelines on the use of cash, for example, reflect increasingly broad acceptance of cash programming, and the agency recently lifted the EURO 100,000 ceiling on cash programmes (DG ECHO, 2009). In the UK, the Humanitarian Emergency Response Review recommended that DFID 'should… make cash based responses the usual relief and recovery position for its partners' (Ashdown, 2011: 24). USAID has recently changed its Food For Peace Title II policy to include cash transfer programming explicitly.

Such policy support at the donor level has led to a marked increase in funding for cash programming in humanitarian operations: Development Initiatives report that spending on cash transfer programmes rose from $74.9 million in 2006 to $188.2 million in 2010 (Development Initiatives, 2012).

Although funding subsequently fell in 2011, the general trend appears to be an increase in the use of cash, not least because large agencies plan to increase their activities in this area significantly (Haley Bowcock, personal communication, December 2012). As a programming approach, 'cash-based work in humanitarian relief has shifted… from radical and risky to… mainstream' (Ramalingam et al., 2009: 43).

To what degree has evidence played a role in this policy shift? The key points of the ALNAP Study documenting the evolution of the use of cash in humanitarian programming are worth repeating here (Ramalingam et al., 2009a). The study suggested that although there was a fairly long history of using cash in emergency response, it was not until 2000 that these scattered experiences were reviewed methodically in a single document: *Buying Power: the use of cash transfers in emergencies* (Peppiatt et al., 2000). This was followed by work from the Humanitarian Policy Group (HPG) at ODI, which published a series of papers considering the utility of cash in emergency contexts. Many of those who were involved in adopting cash programming 'cited the work of… HPG as crucial in… persuading a number of agencies to initiative their own projects'; further, 'credible research documenting the viability of cash in various settings… helped organisations to advocate, internally and externally' (Ramalingam et al., 2009a: 63–64). Research (often in the form of case studies) and evaluations of cash programming continue to be conducted and are collated by the Cash Learning Partnership (CaLP). The CaLP website currently contains 45 evaluations of cash programmes and 40 research reports.

ALNAP's assessment was that 'research and evaluation played an important role' in the acceptance of cash programming (Ramalingam et al., 2009a: 63). Colleagues at CaLP agree on the importance of evidence. They see a lack of evidence in certain areas (particularly around the cost-efficiency and cost-effectiveness of cash as opposed to in-kind assistance) as a constraint to greater acceptance of the approach, and have recently developed a research programme to address some of these evidence gaps.[19] At the same time, there is a general recognition that evidence, on its own, is not sufficient to overcome the doubts and concerns felt by many agencies around the use of cash programming, particularly in complex emergencies. A recent article by Degan Ali argues that, although 'evidence was available that cash transfers were a viable and effective option' in south central Somalia, the 'humanitarian community's aversion to risk made them reluctant to use cash programming at scale early on' and so 'despite a proven history of effectiveness in the region, the [eventual] decision to use cash was more a result of the right personalities and a lack of alternatives than any assessment of the efficacy and appropriateness of cash in meeting basic needs' (Ali, 2012).

19 See www.cashlearning.org/what-we-do/research-focus

The introduction of minimum standards

Buchanan-Smith considers the importance of research evidence in the development of humanitarian policy in a rather different context. Her assessment of 'How the Sphere Project Came into Being' looks at one particular change – the decision to introduce voluntary minimum standards for humanitarian action – and traces the complex relationship between the Joint Evaluation of Emergency Assistance to Rwanda (JEEAR) and the development of the Sphere standards. She concludes that, although the JEEAR had a 'very big impact' (Buchanan-Smith, 2005: 17), and helped to focus attention on the need to establish minimum standards, it was by no means the sole cause of these policy changes, which had 'less to do with research, more to do with growing concern[s] [in the humanitarian sector]' (ibid.: 22). Indeed, while some of the JEEAR research was influential, many of the most important conclusions 'were ducked and have been consistently evaded' (ibid.: 24). We consider one of the conclusions that did not lead to change below.

From the examples of cash programming and minimum standards, it would appear that evidence can – and does – contribute to the development of policy in the humanitarian sector, but that 'the model of policy-making as a rational process that gathers evidence and provides guidance for appropriate actions is highly questionable' (Clarke and Ramalingam, 2008: 32). Evaluations, for example, are 'important, but only one of the resources and influences for change. [They are] generally given a middle ranking in terms of… value to decision maker' (Sandison, 2006: 3).

Policy development is not exclusively evidence-based, and evidence is not always used to develop policy. Our final two examples illustrate the failure of the humanitarian system to make strategic or policy changes on the basis of evidence: the consideration of cultural factors in humanitarian programming; and early intervention in drought-related emergencies.

“Policy development is not exclusively evidence-based, and evidence is not always used to develop policy.”

The consideration of cultural factors in humanitarian programming

A failure to consider cultural factors is one of the conclusions to emerge from the JEEAR and is an issue that appears to have been 'consistently evaded' according to Buchanan-Smith (2005: 25).

In Rwanda, the evaluation team noted that '[b]y and large, relief agencies had only a very limited understanding of the structure of Rwandese society and very little account had been taken of the views of beneficiaries… a large number of the relief agency personnel had not previously worked in the region, knew little about Rwandese society and, as a result, were oblivious to many of the issues of concern to the ordinary, Kinyarwanda-speaking Rwandese' (Borton et al., 1996a: 176).

This lack of contextual knowledge led to a series of mistakes that decreased the effectiveness, efficiency and *relevance* of the response, including: distribution of inappropriate commodities; distribution of commodities through commune-based mechanisms that excluded vulnerable people and allowed officials to build a power base that contributed to insecurity; and support to a policy of early repatriation.

Recognising that many of these mistakes were not made by NGOs that had experience of working in Rwanda, the evaluation concluded that it was 'imperative that NGOs operating in complex emergencies: field qualified professional staff with previous work experience in such settings and appreciation of the need to be sensitive to the local culture; establish partnership with local organisations [and]; include at least some staff or advisors with considerable experience in the country' (Borton et al., 1996b: 61).

Over the next decade, these findings were echoed in academic research that pointed to the importance of understanding the local context in which an emergency response was taking place and of taking the perceptions of local people into account (Dijkzeul, 2010). Evaluations have returned regularly to this theme (see Ali et al., 2005; Oxfam, 2004; World Vision, 2011; Nicholson and Desta, 2010; Tache, 2010).

However, over a decade after the Rwanda response, in 2006, the synthesis report of the Tsunami Evaluation Coalition still reported widespread 'brushing aside…[of]local organisations;… displacement of able local staff by poorly prepared internationals; dominance of English as a 'lingua franca';… applying more demanding conditions to national and local 'partners' than those accepted by international organisations;… and poor-quality beneficiary participation' all of which led to 'inequities, gender and conflict-insensitive programmes, indignities, cultural offence and waste' (Telford and Cosgrave: Synthesis, 2006: 93–4).

The situation does not seem to have greatly improved since the tsunami response. Sixteen years after the JEEAR, and despite evaluative and research evidence that suggests a need for change, beneficiaries still feel inadequately consulted (ALNAP, 2012); very few 'national' staff are promoted to senior operational positions (Buchanan-Smith and Scriven, 2011); local NGOs and civil society organisations are often marginalised in relief operations (ALNAP, 2012); and international staff turnover remains high, preventing decision-makers from obtaining any in-depth knowledge of the social, economic and political context in which they are working (Bhattacharjee and Lossio, 2011; Currion, 2010; Darcy, 2012).

Early intervention in drought-related emergencies

This is not the only failure to create robust policy responses to situations where evidence suggests that change is required. Our final example concerns the way in which the international humanitarian system responds to drought in pastoralist areas.

The 'traditional' humanitarian response to drought has been one of large-scale food distributions, triggered in general by unacceptably high levels of malnutrition. Over the past two decades, however, there have been calls to move to an alternative 'early response' model, in which agencies respond to early warning of drought with a series of livelihood interventions, supporting the health of pastoralist herds and maximising income from livestock sales.

Catley and Cullis, in their paper *Money to burn* (2012), note that these approaches had been used in the Sahel and the Horn of Africa in the 1980s and 1990s. In 2001, Aklilu and Wakesa, reflecting on the 1999–2001 drought response in Kenya, concluded that 'the policy framework of drought response needs to be rethought… moving beyond food relief… to support and maintain, not the people themselves

but their capacity to trade and support their livestock' (2001: 33). Four years later, Jaspars reached similar conclusions, having reviewed the literature and conducted case studies (Jaspars, 2006).

Over the rest of the decade a series of other evaluations and research documents lent support to the idea that a significant policy shift was required (see Sadler et al., 2009; VSF, 2009; Burns et al., 2008; ODI, 2006). The evidence suggested that early response was more effective, more acceptable to local populations, and significantly more cost-efficient (Abebe et al., 2008).

These studies, and the programmes on which they were based, led to a limited response. A previous ALNAP paper on the topic notes that some donors have introduced multi-year funding and flexible funding mechanisms, to allow relevant responses to take place without the need to appeal for new funds (Hedlund and Knox Clarke, 2011). However, these initiatives are 'generally small scale and do not match the needs of affected populations' (ibid.: 6).

Over the past decade, livelihoods interventions have been generally under-funded (HPG, 2006); were not prioritised for the UN's Central Emergency Relief Fund (CERF) (Pantuliano and Wakesa, 2008); and in Ethiopia accounted for only 2.2% of total funding for drought relief in 2011 (Catley, 2012). In 2012, 11 years after Aklilu and Wakesa's call to rethink drought response, a DFID-funded report recommended that 'early response and resilience building measures should be the overwhelming priority response to disasters' (Venton et al., 2012). The same report estimated that, had the international community used de-stocking as a default option over that decade, there would have been savings of around $11 billion. More importantly, 'if an early response had saved even a small proportion of… lives [lost as a result of the 2010/11 drought] thousands of children, women and men would still be alive' (Save the Children and Oxfam, 2012: 13).

Tackling the constraints to the use of evidence

There are, of course, many constraints to using evidence to develop humanitarian policy. Both the JEEAR and the TEC noted the very real disincentives to generating any evidence that suggests an agency, intervention, or approach may have 'failed', and without this it becomes difficult to create a solid body of evidence. In many cases, evidence is scattered and is not available in a single, comparable format (Borton et al., 1996a; Redmond, 2010). In the case of early interventions, Levine and colleagues have noted that the number of programmes, and therefore the evidence-base, is still fairly limited (Levine et al., 2011).[20] But lack of comparable evidence is not sufficient explanation for the lack of attention policy-makers have given to limited contextual knowledge or late response.

Another explanation for the relative inaction of the humanitarian community in the face of evidence is that many issues are just too difficult to solve. In other words, there is little to be gained from increasing contextual knowledge where other factors will make it difficult, if not impossible, to put this knowledge to use. These factors

20 Andy Catley suggests another interesting reason why the formal evidence base for early intervention might not be large: that the approach is based on such strong 'causal logic' that practitioners have not felt any requirement to test the assumption formally – which raises the interesting question of the role of logic models in providing evidence.

include: 'the inflexibility and supply-driven nature of the international relief system' (Borton et al., 1996a: 177); 'donor stipulated restrictions on how [agencies] use funds' (TGLLP Steering Committee: 11), as aspect of the 'path dependence' discussed earlier; and 'the urgency to spend money visibly' (Telford and Cosgrave, 2006: 93).

There is probably more than a grain of truth in this. But again, it is only part of the story. Policy changes that aim to build long-term partnerships with local civil-society actors, ensuring more locally recruited staff are in decision-making positions, or reducing turnover in emergencies, seem feasible and would help to ensure programmes are based on stronger contextual knowledge. Donors could release more funds earlier in slow-onset disasters, and agencies could be better prepared to intervene. As the example of cash programming shows, evidence can contribute to policies that challenge existing elements of the humanitarian paradigm – in the case of cash, the perception that 'cash was not feasible because recipients could not be trusted to spend it effectively' (Ramalingam et al., 2009a: 44). So why, so often, does evidence seem to be ignored? And what does it take to get evidence used?

The challenges of using evidence to develop policy are not exclusively practical: Sandison finds that it is particularly difficult to take (instrumental or general) action on evaluations which 'challeng[e] strongly held beliefs and behaviour embedded in the organisation's culture' (Sandison, 2006: 111), and concludes that 'using evaluation is as much a people issue as a technical one' (ibid.: 132). Clarke and Ramalingam, in their study of change in humanitarian organisations, note that 'interviewees talked about "visceral responses" to what were, on the surface, fairly simple technical changes' (2008: 45).

Effective change – including the development and introduction of new policy – requires a process that addresses the rational, political and emotional needs of stakeholders in the organisation. Recognising this, the Research and Policy in Development (RAPID) programme at ODI has created a framework to examine the influence of research on policy (Young and Court, 2004). The framework looks not only at the credibility and communication of the evaluation information, but also at the links between the evaluators, policy-makers and other networks; at the political context; and at the influence of the external environment. These factors also seem to be important in determining the degree to which evidence influences humanitarian policy.

The RAPID framework emphasises the importance of communicating evidence to decision-makers. In the case of cash, the ALNAP Study found that 'using results [of evaluations] in simple and powerful ways … was crucial' (Ramalingam et al., 2009a: 3). As mentioned, although cash programmes had been around for some time, the 2000 British Red Cross report *Buying power* was the first time that their results became readily accessible to policy-makers.

HPG, and latterly CaLP, have been influential in ensuring that evidence and learning are available and collated. Similarly, the JEEAR 'clearly laid out and analysed what most humanitarian agencies already knew to be the case' (Buchanan-Smith, 2005: 22) and benefitted from a funded follow-up process that allowed the evaluation team to 'sell' the report and the key messages that it contained. By contrast, the lack of attention to issues of context may result, in part, from the lack of a clear

synthesis of the evidence: 'while attention to local perceptions of humanitarian action has been increasing, it has not been systematic enough… these studies rarely refer to each other' (Dijkzeul and Wakenge, 2010: 1146). Given the importance of making evidence accessible, the work of agencies such as Oxfam, Care and NORAD in synthesising and publicising research and evaluations (Hallam, 2011), and of groups and networks such as ALNAP can all contribute to a more evidence-based system (Dijkzeul et al., 2012).

It would appear however, that access to information is not sufficient to guarantee its use. Several studies of the humanitarian world have suggested that humanitarian decision-makers at all levels tend to be strongly influenced by the attitudes and opinions of their peers (Clarke and Ramalingam, 2008; ALNAP, 2003; Sandison, 2007; Darcy, 2009; Maxwell et al., 2013). This suggests that knowledge, in the humanitarian sector, is socially constructed and validated, and that for evidence to be used, it first needs to become a part of the humanitarian discourse.

The importance of networks and relationships in making knowledge 'acceptable' has been noted elsewhere. Latour has shown 'how making science is a social endeavour where enrolling people into accepting certain truths depends more on social relations than on the use of scientific methods' (quoted in Hilhorst, 2003) while Jones and Mendizabal suggest that 'direct interpersonal relations between staff and both researchers and evaluators… matter a great deal' in getting evidence used (Jones and Mendizabal, 2010).

"Knowledge, in the humanitarian sector, is socially constructed and validated."

This tendency, which has worrying implications for the ability of 'local' knowledge to influence the direction of humanitarian action, has been influential in the acceptance of cash as a programming tool. Ramalingam et al. (2009a: 55) suggest that 'the emergence of a dispersed group with field based experience who began to explore…the possibilities for cash programming, and address the particular concerns of sceptics', and which led to the cash-based learning initiative following the 2004 Tsunami (and subsequently CaLP), was important in generating acceptance for the approach. This social momentum, combined with evidence from evaluations, remains important in the continuing development of cash programming.[21]

Similarly, the JEEAR was an inter-agency initiative, with broad participation from across the humanitarian system: the research was social in nature from the beginning. In contrast, Levine et al. (2011) suggest that there is no 'platform' for discussion of early response, and that most discussions are bilateral, and relate to specific programmes: the social network around early response does not seem to exist.

The RAPID framework also highlights the importance of (organisational) politics and external pressure in determining the degree to which evidence is used. Humanitarian policy-makers are selective, 'filtering' evidence, and they 'ultimately make the decision about which of the researchers' recommendations for policy change they [are] prepared to accept' (Buchanan-Smith, 2005). As a result, 'the humanitarian system… is most responsive to change under pressure, when the push factors are strong' (ibid.: 98).

21 According to members of the CaLP discussion forum, who kindly provided their views on this topic.

These push factors were particularly strong in the aftermath of the Rwanda response, for example, with agencies engaged in internal debates about how to improve and donors demanding action, and there is no doubt that this ensured that the evidence of the JEEAR teams was at least given a hearing.

Some recommendations – such as those around standards – were then pushed through the filter. In the case of cash programming, developments took place against a background of long-running concern over the effects of food aid, which was influenced externally by a variety of factors. These included the massive increase in funding that took place after the Tsunami, the support of governments in the Indian Ocean region for cash programmes, and the global increase in food and oil prices in 2008, which made food-aid delivery more expensive.

In the case of early response, on the other hand, both organisational and political factors seem to work against policy change. Many agencies may avoid livelihoods programming because they lack the skills and contextual knowledge required (Aklilu and Wakesa, 2001), and because the 'fire brigade' model of establishing a presence in an area when a disaster occurs is not effective for early response (HPG, 2006). In addition, some organisations receive significant funding from monetisation of the (large quantities) of food aid required to address critical conditions (Jaspars, 2006): the relatively limited sums required for livelihood support would not provide the same level of income.

Donors, meanwhile, can be unwilling to respond on the basis of prediction alone, requiring 'hard data' before committing tax-payers' money (Save the Children and Oxfam, 2012; Levine et al., 2011) and are often incentivised to spend larger sums of money than the NGOs' request for early interventions. Levine et al. quote one donor representative as saying 'NGOs take small amounts of money…if we give a large cheque to the UN, we can write it off our books straight away' (2011: 7). Significant constraints to change also exist at the political level: 'National governments often see an emergency declaration as a sign of weakness' (Save the Children and Oxfam, 2012) and so delay 'calling' an emergency until it is too late for livelihoods approaches to be particularly effective (Hedlund and Knox Clarke, 2011).

The path from evidence to policy is seldom straight or clear. In some cases – and particularly those where the evidence challenges received wisdom or standard approaches, it may well peter out altogether. But in many cases, evidence does inform (if not guide) policy: this is particularly the case where the evidence becomes generally accepted, either through the support of coalitions or other platforms, or under the influence of a crisis which makes people more open to considering change.

IN BRIEF:

- There is a difference between the direct 'instrumental' use of evaluations and their longer term, indirect role in influencing humanitarian policy.
- Evaluations are often used instrumentally (for the immediate redesign of the programmes or projects evaluated), although this use is selective.
- 'Hard to implement' recommendations are less likely to be used.
- ALNAP has researched the factors that lead to instrumental use: the most important seems to be a close engagement of the decision-makers with the

evaluation process and many humanitarian organisations are now using methods to support this type of engagement.

- The role of evaluations (and evaluative research) in influencing policy is considered through the use of four examples: cash programming; designing programmes for contextual appropriateness; the evolution of humanitarian standards; and early response to drought situations.
- Evaluations are only one source of evidence for policy development and their influence on policy is often limited by the fact that evaluations can be hard to find (in part a result of a reluctance on the part of agencies to circulate evidence that reflects badly on their programmes), and by the natural resistance to change displayed by individuals and organisations in the humanitarian sector.
- In addition, many evaluations point to changes that would require fundamental rearrangement of humanitarian architecture and the humanitarian 'business model'.
- Evaluations are more likely to inform policy in situations where there is already pressure for change.
- Platforms that bring together a variety of actors, and that 'socialise' the knowledge found in evaluations, are also important in supporting the use of evaluations.

Conclusions and recommendations: how can the quality and use of evidence in humanitarian action be improved?

5.1 Improving the quality of evidence

Good evidence – and particularly evidence on the nature of humanitarian needs and on the success or failure of responses – is an essential component of humanitarian programming. As a result, there are processes for information collection and analysis at several stages in the standard humanitarian 'programme cycle'.

An assessment of the quality of evidence produced by these processes (using the criteria for 'good' evidence outlined in Section 2). suggests that there is significant room for improvement. This review also suggests that improvement is possible: while humanitarian contexts present serious constraints to the generation of evidence, many organisations within the humanitarian system are finding ways to overcome them.

On the basis of this review, we suggest that efforts to improve the quality of evidence in the humanitarian sector should consider five key principles.

ROBUST METHODOLOGY: much of the evidence generated in the sector is obtained through qualitative methods. However, these methods are often poorly understood and implemented. Agencies can – and should – improve their methodological rigour in this area. At the same time, some agencies are experimenting with the use of quantitative methods of data collection and analysis. These initiatives have demonstrated the possibilities, as well as the limitations of these approaches, and should be encouraged. Tentative moves towards 'mixed method' approaches to evidence generation should be welcomed.

All early warning, assessment, evaluation and research reports produced by humanitarian actors should be clear about the strength of evidence used, and contain a description of the methodology used for information collection and analysis, and the limitations of this methodology.

Organisations should share experiences and learning around their use of qualitative and quantitative methodologies, bilaterally and through networks and inter-organisational groups such as ALNAP, ACAPS and the IASC Needs Assessment Task Force.

These networks should synthesise best practices and support the production of training materials and other tools for dissemination of best practices.

Humanitarian organisations should consider establishing bodies to review the quality of their reports.

Humanitarian organisations should assess whether they have the capacity to conduct evidence collection and analysis relevant for their work and, where they do not, they should fill gaps through strategic partnerships, recruitment and training.

Humanitarian actors should partner with, and learn from, academics to improve the standard of evidence generation.

EFFICIENCY: evidence is an input to humanitarian programming, and as such the time and money invested in generating evidence should be considered in the context of overall expenditure on current and future programmes.

Donors and humanitarian organisations should consider potential benefits explicitly – both direct and longer term – when investing in evidence collection, and be prepared to invest sums proportionate to the savings or improvements that could be made.

COLLABORATION: one of the key constraints to the effective generation (and use) of high quality evidence in the humanitarian system is a lack of collaboration within and between organisations. Evidence is often collected in functional or organisational silos. This prevents rigorous cross-checking and quality control, encourages duplication of activities, and effectively denies most humanitarian actors access to large quantities of evidence. Initiatives that aim to generate humanitarian evidence should recognise a responsibility to contribute to a sector-wide 'body of evidence'.

Donors should use inter-agency networks and platforms to identify the 'evidence gaps' that exist within the humanitarian system, and develop common approaches to filling them.

All humanitarian actors, and particularly donors, should continue to support processes to identify global standardised frameworks that can be used for joint or coordinated early warning, assessment, and evaluation.

HCTs, Clusters and Inter Cluster Coordination Mechanisms should clarify evidence needs at country level and encourage joint or coordinated assessments and evaluations.

All evidence collection activities should include budgets and plans to ensure the dissemination of evidence, and maximise the use of system-wide bodies (such as ACAPS, ALNAP and CaLP) to support dissemination.

CONTINUITY: short funding cycles and rapid staff rotation tend to work against the collection of bodies of comparable evidence over time. As a result, evidence is often produced episodically, in one-off assessments and evaluations. This significantly reduces our ability to understand the effects of humanitarian intervention over time, and so to be able to say, with confidence, when and how we should intervene in the future. Humanitarian actors should see evidence generation as a process, and not an event, and aim to build the body of evidence over time.

Humanitarians should refrain from the continual redesign of assessment and monitoring systems. Assessment systems and monitoring systems should be designed to complement one another. Where systems are redesigned, care should be taken to continue to collect data on key indicators.

Humanitarians should support national work to create monitoring and other systems that can provide baseline data, and longer time-series of data for key indicators, through disaster risk reduction and contingency planning.

INCLUSION AND OWNERSHIP: the people who are concerned most intimately with humanitarian action – those who are personally affected by a crisis – tend to be marginalised by current approaches to evidence generation in two important ways. First, they do not get to ask the questions: evidence is generally collected to meet the needs of international organisations, rather than those of affected people. Assessments, for example, tend to obtain evidence for what sort of aid is needed, rather than for whether aid is needed in the first place. Second, their answers are often not seen to be important: this review has suggested that even when gathering evidence to understand whether a situation is a crisis, or whether a particular intervention 'worked', the knowledge and opinions of those most directly affected tend to receive only limited attention. In future, international organisations should be clear about why they are collecting evidence and who it is for; should consider the degree to which they can collect evidence that is of use to civil society organisations in affected areas, and should make strenuous efforts to include the voices of affected people in their evidence collection.

International organisations should identify the information and evidence needs of civil society organisations in disaster-affected areas and provide information to meet these needs. Where relevant, they should also provide training and resources to allow national and local organisations to collect information.

Early warning, assessment, evaluation and research reports produced by humanitarian actors should present local knowledge and proposals clearly. Where recommendations are not in line with this local information or proposals, the authors should explain why this is the case.

Information should, in all cases, be collected on the basis of informed consent. In addition, organisations collecting information should identify ways to inform people about how the information was used, and about any decisions that were made on the basis of this information.

5.2 Improving the use of evidence in humanitarian decision-making

This review also considered the degree to which decision-makers within the humanitarian system use evidence when taking decisions. In general, we found that there is room for significant improvement in the use of evidence by humanitarians: there are many instances where decision-making processes have ignored important evidence from early warning systems, needs assessments or evaluations, with negative consequences for the lives and livelihoods of disaster-affected people.

This is particularly the case where the evidence challenges preconceived interpretations of a situation, or points to a need for radical change.

At the same time, we should not be overly simplistic about how evidence can or should be used in decision-making. Most decisions in the humanitarian sector require consideration of multiple bodies of evidence, often collected using different methodologies, and sometimes indicating different courses of action. In these circumstances, it is not realistic or desirable to hope that the evidence alone will determine the decision. There will always be a necessary element of judgement in decision-making, and we should aim to ensure that these judgements are, in all cases, informed by the best evidence available.

Improving the use of evidence – collection, analysis and dissemination

On the basis of this review, we suggest that evidence is more likely to influence decision-making where those who are responsible for generating evidence are guided by the following principles:

ACCESSIBILITY: decision-makers will seldom have the time (and on occasion may lack the inclination) to search through databases, reports or spreadsheets for the information they need to make a robust decision. Evidence should be packaged in ways that make it easily accessible: short reports in jargon-free language, infographics, and searchable 'portals'.

Humanitarian learning networks (such as ALNAP), inter-agency initiatives (such as the global clusters), and specialised organisations should endeavour to ensure that information and evidence can readily be accessed, through:

- synthesis of best practices and lessons learned from evaluations and research
- the creation and maintenance of easily accessible, searchable, platforms for documentary and other forms of evidence, with those entities creating databases ensuring that they collaborate with existing initiatives to prevent duplication and confusion.

TIMELINESS: information that arrives after decisions have been made will often be ignored, and the most important decisions – the ones that set the broad scope of a programme – will often be made early on. Groups or individuals producing evidence should aim to understand the decision-making process and to make relevant evidence available at critical points in the process.

Humanitarian organisations responsible for early warning, assessment and evaluation should assess the degree to which these products are being produced at the right time to influence decisions.

Where necessary, they should attempt to alter their processes to ensure that the information is available in a timely manner.

BROAD CIRCULATION THROUGH A VARIETY OF MEDIA: knowledge in the humanitarian sector appears to be, in large part, socially constructed. Decision-makers are influenced by what their peers know and believe. As a result, information that has been widely circulated, and is widely known and discussed is far more likely to be used as evidence to support or oppose a certain course of action.

Organisations producing evidence should ensure that it is disseminated actively through a variety of methods including webinars, communities of practice, meetings, videos and training courses.

They should aim to engage a variety of groups, including policy-makers, practitioners and organisational leadership, through these various media.

Donors supporting evidence generation should ensure that budgets include plans for dissemination of results.

Improving the use of evidence – decision-making

While humanitarians have a responsibility to ensure that their evidence is accessible, timely and widely circulated, these activities alone will not be enough to increase the engagement of decision- makers with the evidence. ALNAP member organisations should also take action to improve their approach to decision-making.

ESTABLISH CLEAR DECISION-MAKING PROCESSES: evidence is more likely to be used where organisations institute decision-making processes that include clear steps that require evidence to be considered.

Humanitarian organisations should clearly outline the process by which programmes are designed (and redesigned), indicating key stages at which information is required, and – where possible – the nature of the information required to make decisions.

Donors should require that the humanitarian organisations that bring them funding proposals provide the best evidence available (considering time and resource implications) to demonstrate that:

- external assistance is needed
- the proposed response is the best option available
- the proposing organisation has the capacity to implement the response effectively.

Humanitarian agencies should consider introducing response/decision tracking matrices (where these are not used already) to allow stakeholders to see the degree to which the organisation has acted on evaluation recommendations

In the field of early warning/early response, the humanitarian system should establish a common platform to consider the entire process from early warning to early action, focussing on the role of early warning systems in informing decisions; the location of decision-making authority in response agencies; and joint learning around early response activities.

ENSURE THAT THERE ARE INCENTIVES FOR THE USE OF EVIDENCE: using evidence in decision-making will often incur a time cost (in finding and processing the evidence) and may also lead to decisions that threaten previously established ways of working, or that are otherwise 'hard to implement'. It may be easier for decision-makers to avoid looking for, or using, evidence. To address this, organisations should consider incentives for the use of evidence.

Any future standards and associated certification processes used in the humanitarian sector should require agencies to meet agreed standards around the quality of evidence collection.

Humanitarian organisations should clarify decision-making roles and expectations of decision-makers. It should be clear who is responsible for making which decisions: those responsible should be aware of the organisation's expectations around use of evidence.

Box 6. What will ALNAP do?

We hope that all ALNAP member organisations will consider and implement these recommendations. Meanwhile, on behalf of the Network, the ALNAP Secretariat will:

- work with the Evaluation Community of Practice to develop and share practical guidance on ways to improve the evidential quality of humanitarian evaluations
- develop training materials to support ALNAP Members in improving the evidential quality of evaluations, based on the *Evaluating Humanitarian Action (EHA)* guide (Buchanan-Smith and Cosgrave, 2013).
- continue to update and refine the resources library: ensuring that it contains the largest possible number of evaluations, lessons documents and reports on humanitarian action, and that it is accessible and easy to search.
- continue to experiment with a variety of research methodologies (including: structured literature reviews; pattern-matching approaches to case study analysis; quantitative survey methods) to ensure that ALNAP reports are based on the most robust evidence possible: publish and disseminate our experiences with the use of these methodologies.
- Work with ALNAP Members (through our work on operational leadership) to better understand decision-making processes and risk/uncertainty avoidance, and publish concrete recommendations to improve decision-making based on robust research.

Annex 1: Methodology

The initial version of this report, *Evidence and Knowledge in Humanitarian Action*, was based on a literature review, augmented by a small number of interviews. It was circulated as the background document for the ALNAP 28th Annual Meeting and presented to the meeting on 5 March 2013.

The current report is based on that background document, which has been rewritten to incorporate:

- comments made by the panel and meeting participants on the background paper itself
- recommendations made by meeting participants in the final session of the meeting on 6 March 2013, and
- information from the 26 presentations made at the meeting.

The presentations that have been incorporated into this report were selected on the basis of an open call to humanitarian practitioners and academics, through the ALNAP Membership. Presenters were asked to discuss 'new learning and emerging best practice in the understanding of evidence and in the collection, analysis and use of evidence in humanitarian action' (ALNAP, 2013: 1). In all, 26 presentations were selected from 110 proposals. All presentations were made in English.

This paper also includes results of a further review of the literature, informed by topics that emerged at the Annual Meeting. The paper has been peer reviewed by a panel of experts.

The key limitations in this approach relate to selection bias. The literature reviews were not based on a structured search of academic databases, and did not use formal inclusion or exclusion criteria (Hagen-zanker and Mallett, 2013).

Similarly the choice of presentations included in the Annual Meeting and referenced here, while less open to influence by a single author, is biased by the channel of selection (through the ALNAP Network) and by the fact that all presentations were given in English.

Annex 2: List of interviewees and contributors

The following individuals were kind enough to agree to discuss issues in this report in person or by phone, or answered questions over email:

Miriam Ait Aissa, ACF

Ben Allen, ACF

Haley Bowcock, CaLP

Jeff Crisp, UNHCR

Andrew Catley, Tufts University

Mary Dixon-Woods, University of Leicester

Merry Fitzpatrick, World Concern

Richard Garfield, CDC

Francois Grünewald, Groupe URD

Joyce Luma, WFP

John Mitchell, ALNAP

Lili Mohiddin, CaLP

Lars Peter Nissen, ACAPS

Jock Paul, UNOCHA

Chloe Puett, Tufts University

Anthony Redmond, University of Manchester

Cecile Salpeteur, ACF

Louise Shaxson, ODI

Sharon Truelove, British Red Cross

Vivien Margaret Walden, Oxfam

Bibliography

Abebe, D. et al. (2008) 'Impact of a commercial destocking relief intervention in Moyale district, southern Ethiopia', in *Disasters* 32(2): 167–189. www.alnap.org/resource/8200.aspx

ACAPS (2010) 'Rapid Initial Needs Assessment for Haiti (RINAH)'. Port-au-Prince: ACAPS. www.alnap.org/resource/9811.aspx

ACAPS (2011) 'Joint Rapid Assessment of the Northern Governorates of Yemen'. Sanaa: ACAPS. www.alnap.org/resource/9812.aspx

ACAPS (2012) 'Technical Brief: Estimation of affected population figures'. Geneva: ACAPS. www.alnap.org/resource/9813.aspx

ACAPS (2012) 'Coordinated Assessments in emergencies. What we know now : Key lessons from field experience'. Geneva: ACAPS. www.alnap.org/resource/7912.aspx

ACF (2011) 'Learning review'. London: ACF. www.alnap.org/resource/7001.aspx

Aklilu, Y. and Wekesa, M. (2001) 'Livestock and Livelihoods in Emergencies: Lessons Learnt from the 1999–2001 Emergency Response in the Pastoral Sector in Kenya'. Nairobi: OUA IBAR. www.alnap.org/resource/9815.aspx

Alexander, J. et al. (2013) '2013 Humanitarian accountability report'. Geneva: HAP International. www.alnap.org/resource/8758.aspx

Ali et al. (2005) 'Cash Relief in a Contested Area: Lessons from Somalia: HPN Network paper no. 50'. London: HPN/ODI. www.alnap.org/resource/7585.aspx

Ali, D. (2012) 'A Deadly Delay: Risk Aversion and Cash in the 2011 Somalia Famine,' in Bailey,S. and Ridsdel, B. (eds), New learning in cash transfer programming [special feature] *Humanitarian Exchange 54*. London: HPN/ODI. 10–15. www.alnap.org/resource/9817.aspx

Allen, B. (2012) 'Collecting and Reporting best practices from the field'. *Msg 1, thread 2 under Capacity Area 2: Evaluation policy and purpose,* in *ALNAP Evaluation capacities – Community of Practice* (internal forum) London: ALNAP. (Accessed: 9 September 2012) www.alnap.org/resource/9818.aspx

ALNAP (2003) 'Learning by Field Level Workers,' in *ALNAP review of Humanitarian Action in 2003.* London: ALNAP. www.alnap.org/resource/5208.aspx

ALNAP (2010) 'The State of the Humanitarian System – Assessing Performance and Progress: a Pilot Study'. London: ALNAP/ODI. www.alnap.org/resource/8746.aspx

ALNAP (2012) 'The State of the Humanitarian System 2012 edition'. London: ALNAP/ODI. www.alnap.org/sohsreport

ALNAP (2013) 'Evidence and Knowledge in Humanitarian Action – Call for Presentations'. London: ALNAP/ODI. www.alnap.org/resource/7162.aspx

Anderson et al. (2012) 'Time to listen: Hearing people on the receiving end of international aid'. Cambridge, Mass: CDA Collaborative Learning Projects. www.alnap.org/resource/8530.aspx

Argren, R. (2013) 'What goes up must come down: the challenges of getting evidence back to the ground'. Presented at, *Evidence and Knowledge in Humanitarian Action: ALNAP 28th Annual Meeting*, Washington, 5–7 March. www.alnap.org/resource/8015.aspx

Ashdown, P. (2011) 'Humanitarian Emergency Response Review'. London: HERR. www.alnap.org/resource/6355.aspx

Assessment Working Group for Northern Syria (2013) 'Joint Rapid Assessment of Northern Syria – Aleppo City Assessment'. Syria: Assessment Working Group for Northern Syria. www.alnap.org/resource/8755.aspx

Avila et al. (2010) 'Global mapping of technology for transparency and accountability'. London: Open Society Foundation. www.alnap.org/resource/6421.aspx

Banatvala, N. and Zwi, A. (2000) 'Public health and humanitarian interventions: developing the evidence base', in *The British Medical Journal* 321:101–5. www.alnap.org/resource/9820.aspx

Beck, T. (2006) 'Evaluating humanitarian action using the OECD-DAC criteria – An ALNAP guide for humanitarian agencies'. London: ALNAP/ODI. www.alnap.org/resource/5253.aspx

Bekele, G. and Abera, T. (2008) 'Livelihoods-based Drought Response in Ethiopia: Impact Assessment of Livestock Feed Supplementation', Pastoralist Livelihoods Initiative. www.alnap.org/resource/9936.aspx

Bekele, G. (2010) 'Review of Save the Children US Livestock Marketing Initiative in Southern Ethiopia'. Westport: Save the Children USA. www.alnap.org/resource/9937.aspx

Beynon et al. (2012) 'What difference does a policy brief make?'. Brighton: IDS and 3iE. www.alnap.org/resource/9938.aspx

Bhattacharjee, A. and Lossio, R. (2011) 'Evaluation of OCHA Response to the Haiti Earthquake Final Report'. New York: UN OCHA. www.alnap.org/resource/6002.aspx

Bickman, L. and Reich, S. (2009) 'Randomised Controlled trials: A gold standard with feet of clay?', in Donaldson et al. (2009) *What counts as credible evidence in applied research and evaluation practice.* Thousand Oaks: Sage. www.alnap.org/resource/10028.aspx

Bond. (n.d.). 'The NGO Evidence Principles'. London: Bond. www.alnap.org/resource/9941.aspx

Bonino, F. with Jean, I. and Knox Clarke, P. (forthcoming) Effective feedback mechanisms in humanitarian contexts - Guidance for practitioners and policy-makers based on findings from ALNAP-CDA case studies in Sudan, Pakistan and Haiti. ALNAP Study. London: ALNAP/ODI. www.alnap.org/resource/9944.aspx

Borton et al. (1996a) 'The International Response to Conflict and Genocide: Lessons from the Rwanda Experience Study 3: Humanitarian Aid and Effects'. London: ODI. www.alnap.org/resource/9976.aspx

Borton, J. et al. (1996b) 'The International Response to Conflict and Genocide: Lessons from the Rwanda Experience: Synthesis Report'. London: ODI. www.alnap.org/resource/2517.aspx

Bradt, D. A. (2009) 'Evidence-based decision-making in humanitarian assistance'. *HPN Network Paper* 67. London: HPN/ODI. www.alnap.org/resource/7503.aspx

Broadbent, E. (2012) 'Politics of research-based policy in African policy debates: Synthesis of case study findings'. London: ebpdn. www.alnap.org/resource/9942.aspx

Buchanan-Smith, M. and Davies, S. (1995) 'Famine Early Warning Systems and Response: the Missing Link'. London: IT Publications. www.alnap.org/resource/9943.aspx

Buchanan-Smith, M. (2000) 'Role of Early Warning Systems in Decision-Making Processes'. London: HPN/ODI. www.alnap.org/resource/9946.aspx

Buchanan-Smith, M. and Cosgrave, J. (2013) 'Evaluation of humanitarian action guide: Pilot guide'. London: ALNAP/ODI. www.alnap.org/www.alnap.org/eha

Buchanan-Smith, M. and Beck, T. (2008) 'Joint evaluations coming of age? The quality and future scope of joint evaluations', in *ALNAP Review of humanitarian action: Chapter three*. London: ALNAP/ODI. www.alnap.org/resource/5232.aspx

Buchanan-Smith, M. and Scriven, K. (2011) 'Leadership in Action: Leading Effectively in Humanitarian Operations'. London: ALNAP/ODI. www.alnap.org/resource/6118.aspx

Buchanan-Smith, M. (2005) 'How the Sphere Project Came into Being: A Case Study of Policy Making in the Humanitarian-aid Sector and the Relative Influence of Research' in Court et al. (eds) *Bridging Research and Policy in Development: Evidence and the Change Process.* London: ODI. www.alnap.org/resource/9947.aspx

Burns et al. (2008) 'Impact Assessment of the Pastoralist Survival and Recovery Project, Dakoro'. Niger: Lutheran World Federation and Feinstein International Center. www.alnap.org/resource/5722.aspx

CaLP (2013) 'Protecting Beneficiary Privacy: Principles and Operational Standards for the Secure Use of Personal Data in Cash and E-Transfer Programmes' Oxford: CaLP. www.alnap.org/resource/9750.aspx

Carr, H. (2013) 'Tracking Beneficiaries and Services Delivered'. Presented at, *Evidence and Knowledge in Humanitarian Action: ALNAP 28th Annual Meeting*, Washington, 5–7 March. Washington: Catholic Relief Services. www.alnap.org/resource/8023.aspx

Catley, A. (2013) 'Developing international standards and guidelines with limited hard evidence'. Presented at, *Evidence and Knowledge in Humanitarian Action: ALNAP 28th Annual Meeting*, Washington, 5–7 March. Washington: Tufts University. www.alnap.org/resource/8017.aspx

Catley et al. (n.d.) 'Participatory impact assessment: A guide for Practitioners'. Boston: Feinstein International Centre and Tufts University. www.alnap.org/resource/8094.aspx

Catley, A. and Cullis, A. (2012) 'Money to burn? Comparing the costs and benefits of drought responses in pastoralist areas of Ethiopia', in *The Journal of Humanitarian Assistance*. Ethiopia: FAO. www.alnap.org/resource/9948.aspx

Clarke, P. and Ramalingam, B. (2008) 'Organisational change in the humanitarian sector' in *ALNAP 7th Review of Humanitarian Action*. London: ALNAP/ODI. www.alnap.org/resource/5235

Cooley, A. and Ron, J. (2002) 'The NGO scramble: organizational insecurity and the political economy of transnational action', in *International Security* 27(1): 5–39. www.alnap.org/resource/9949.aspx

Cosgrave, J. (2007) 'Tsunami Evaluation Coalition, Synthesis Report: Expanded Summary, Joint evaluation of the international response to the Indian Ocean tsunami'. London: ALNAP and TEC/ODI. www.alnap.org/resource/5536.aspx

Coyle, D. and Meier, P. (2009) 'New Technologies in Emergencies and Conflicts: The Role of Information and Social Networks'. Washington: United Nations Foundation. www.alnap.org/resource/6572.aspx

Currion, P. (2010) 'Coordination at the Crossroads: NGO coordination in Southern Sudan 2007–2011'. London: ALNAP and TEC/ODI. www.alnap.org/resource/5781.aspx

Damerell, J. (2013) 'The role of standards for evidence based humanitarian response', Presentation at *Evidence in Humanitarian Action, ALNAP 28th Annual Meeting*. Washington: ALNAP. www.alnap.org/resource/8016.aspx

DARA (2011) 'The Humanitarian Response Index 2011'. Madrid: DARA. www.alnap.org/resource/6243.aspx

Darcy, J. and Hofmann, C. (2003) 'According to need? Needs assessment and decision making in the humanitarian sector'. London: HPG/ODI. www.alnap.org/resource/9384.aspx

Darcy, J. (2009) 'Humanitarian diagnostics: the use of information and analysis in crisis response decisions'. London: HPG/ODI. www.alnap.org/resource/9485.aspx

Darcy, J. and Garfield, R. (2011) 'Review of the Rapid Initial Needs Assessment for Haiti', paper prepared for ACAPS. Geneva: Assessment Capacities Project. www.alnap.org/resource/9951.aspx

Darcy et al. (2013) 'The Use of Evidence in Humanitarian Decision Making ACAPS Operational Learning Paper'. Boston: Feinstein International Center and Tufts University. www.alnap.org/resource/8003.aspx

Development Initiatives (2012) 'Global Humanitarian Assistance: Tracking spending on cash transfer programming in a humanitarian context'. Wells: Development Initiatives. www.alnap.org/resource/9952.aspx

de Ville de Goyet, C. and Moriniere, L. C. (2006) 'The Role of Needs of Assessment in the Tsunami Response'. London: ALNAP and TEC/ODI. www.alnap.org/resource/3531.aspx

DFID (2012) 'Promoting innovation and evidence-based approaches to building resilience and responding to humanitarian crises: a DFID strategy paper'. London: DFID. www.alnap.org/resource/9823.aspx

DFID (2013) 'How to note – Assessing the Strength of Evidence'. London: DFID. www.alnap.org/resource/8402.aspx

DG ECHO (2009) 'The Use of Cash and Vouchers in Humanitarian Crises: DG ECHO funding guidelines'. Brussels: DG ECHO. www.alnap.org/resource/9824.aspx

Dijkzeul, D. and Wakenge, C. (2010) 'Doing good, but looking bad? Local perceptions of two humanitarian organisations in eastern Democratic Republic of the Congo', in *Disasters* 34(4): 1139–70. www.alnap.org/resource/9953.aspx

Dijkzeul et al. (2013) 'Introduction to Evidence based action in humanitarian crises'. [Manuscript submitted for publication], in *Disasters*, 37: S1–19. www.alnap.org/resource/9954.aspx

ECOSOC (2007) 'UNICEF evaluation policy'. New York: UNICEF. www.alnap.org/resource/10037.aspx

Epstein et al. (2004) 'Communicating evidence for participatory decision making' in *The Journal of the American Medical Association*, 291(19): 2359–66. www.alnap.org/resource/9955.aspx

Evans et al. (2011) 'Testing treatments: better research for better health care'. London: Pinter and Martin. www.alnap.org/resource/9956.aspx

FAO and WFP (2013) 'Joint Evaluation of Food Security Cluster Coordination'. Rome: FAO/WFP. www.alnap.org/resource/9957.aspx

Fearon et al. (2008) 'Community Driven Reconstruction in Lofa County: Impact assessment'. New York: International Rescue Committee. www.alnap.org/resource/8192.aspx

Featherstone, A. (2011) 'Strength in numbers: a global mapping review of NGO engagement in coordinated assessments'. ECB Project. www.alnap.org/resource/9958.aspx

Featherstone, A. (2013) 'Improving Impact: Do Accountability Mechanisms Deliver Results?'. London: Christian Aid, Save the Children for the Humanitarian Accountability Partnership. www.alnap.org/resource/8388.aspx

Feinstein International Center (2007) 'Impact Assessments of Livelihoods-based Drought Interventions' in Moyale and Dire Woredas, Ethiopia: A Pastoralist Livelihoods Initiative report'. Boston: Feinstein International Center, Tufts University www.alnap.org/resource/3490.aspx

Fenton et al. (eds) (2011) Humanitarian Accountability [special feature] *Humanitarian Exchange* 52. London: HPN/ODI. www.alnap.org/resource/6256.aspx

Fetouh et al. (2013) 'Data quality in remote monitoring – a comparative analysis'. Presented at, *Evidence and Knowledge in Humanitarian Action: ALNAP 28th Annual Meeting*, Washington, 5–7 March. www.alnap.org/resource/8027.aspx

Few, R. et al. (2013) 'Contribution to Change: An approach to evaluating the role of intervention in disaster recovery'. Oxford: Oxfam. www.alnap.org/resource/10305.aspx

Foresti, M. (2007) 'A Comparative Study of Evaluation Policies and Practices in Development Agencies'. London: ODI. www.alnap.org/resource/7773.aspx

Garfield, R. (2013) 'Evidence and needs assessments', Presented at, *Evidence and Knowledge in Humanitarian Action: ALNAP 28th Annual Meeting*, Washington, 5–7 March. www.alnap.org/resource/9959.aspx

Garfield et al. (2011) 'Common Needs Assessments and Humanitarian Action'. *Humanitarian Network Paper* Number 69. London: HPN/ODI. www.alnap.org/resource/6262.aspx

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

Gilman, D and Noyes, A. (2013) 'Humanitarianism in the Network Age'. New York: OCHA. www.alnap.org/resource/9157.aspx

Goldacre, B. (2012) *Bad pharma: how drug companies mislead doctors and harm patients*. London: Fourth Estate. www.alnap.org/resource/9960.aspx

Gottwald, M. (2010) 'Competing in the humanitarian marketplace: UNHCR's organizational culture and decision-making process'. Geneva: UNHCR Policy Development and Evaluation Service. www.alnap.org/resource/9961.aspx

Grünewald, F. (2012) 'Sustaining learning from evaluation: stakeholder engagement and 'customer care'?'. Msg 1, thread 2 under Capacity Area 3: Evaluation processes and systems. Message posted in *ALNAP Evaluation capacities – Community of Practice* (internal forum). London: ALNAP. www.alnap.org/resource/9818.aspx

Grünewald et al. (2010) 'Inter-agency real-time evaluation in Haiti: 3 months after the earthquake'. Plaisans: Groupe URD. www.alnap.org/resource/9962.aspx

Guenther et al. (2010) 'The politics of evaluation: evidence-based policy or policy-based evidence?' Paper presented to *the NARU Public Seminar Series,* 30 November 2010, Darwin. www.alnap.org/resource/9963.aspx

Guerrero, S. (2012) 'Our vision of linking evaluations and learning: ACF Learning Review 2011' *ALNAP blog*. 18th May. London: ALNAP. www.alnap.org/resource/9964.aspx

Guerrero et al. (2013) '2012 Learning review'. London: ACF. www.alnap.org/resource/8219.aspx

Hagen-zanker, J. and Mallett, R. (2013) 'How to do a rigorous, evidence-focused literature review in international development'. London: ODI. www.alnap.org/resource/8675.aspx

Hallam, A. (2011) 'Harnessing the power of evaluation in humanitarian action'. London: ALNAP/ODI. www.alnap.org/resource/6123.aspx

Hallam, A. and Bonino, F. (2013) Using Evaluation for a Change: Insights from humanitarian practitioners. ALNAP Study. London: ALNAP/ODI. www.alnap.org/resource/8980.aspx

Hammersley, M. (1990) *Reading ethnographic research: A critical guide*. London: Longman. www.alnap.org/resource/9967.aspx

Hammersley, M. (2005) 'Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers' case for research-based policy making and practice'. *Evidence Policy* 1(1): 85–100. www.alnap.org/resource/9968.aspx

HAP (2010) 'The 2010 HAP Standard in Accountability and Quality Management'. Geneva: HAP International. www.alnap.org/resource/8125.aspx

Harragin, S., and Chol, C. (1998) 'The southern Sudan vulnerability study'. Nairobi: Save the Children and South Sudan Programme. www.alnap.org/resource/9969.aspx

Hedlund, K. and Knox-Clarke, P. (2011) 'Humanitarian action in drought-related emergencies'. ALNAP Lessons Paper. London: ALNAP/ODI. www.alnap.org/resource/6156.aspx

Henry, G. T. and Mark, M. (2003) 'Beyond use: understanding evaluation's influence on attitudes and actions' in *American Journal of Evaluation* 24(3): 293–314. www.alnap.org/resource/9970.aspx

Hilhorst, D. (2003) 'Responding to disasters: diversity of bureaucrats, technocrats and local people' in *International Journal of Mass Emergencies and Disasters* 21(3): 37–55.www.alnap.org/resource/9971.aspx

Hillier, D. and B. Dempsey (2012) 'A Dangerous Delay: The cost of late response to early warnings in the 2011 drought in the Horn of Africa'. Oxford: Oxfam. www.alnap.org/resource/6291.aspx

House, S. (2012) 'Evaluation of Pakistan flood response 2011/12: using Oxfam GB's global humanitarian indicator tool'. Oxford: Oxfam. www.alnap.org/resource/8366.aspx

HPG (2006) 'Saving lives through livelihoods: critical gaps in the response to the drought in the Greater Horn of Africa: HPG Briefing Paper'. London: HPG/ODI. www.alnap.org/resource/8656.aspx

Humphreys, M. and Weinstein, J. M. (2009) 'Field experiments and the political economy of development' in *Annual Review of Political Science* 12(1): 367–78. www.alnap.org/resource/9972.aspx

Huybregts et al (2012) 'The effect of adding ready-to-use supplementary food to a general food distribution on child nutritional status and morbidity: a cluster-randomized controlled trial'. *PLoS medicine* 9(9). www.alnap.org/resource/9973.aspx

IASC (2011) 'Humanitarian System-Wide Emergency Activation: definition and procedures'. Geneva: IASC. www.alnap.org/resource/8502.aspx

IASC (2012a) 'Multi Cluster/Sector Initial Rapid Assessment (MIRA) Provisional version'. Geneva: IASC. www.alnap.org/resource/9974.aspx

IASC (2012b) 'Operational Guidance for Coordinated Assessments in Humanitarian Crises – Provisional Version'. Geneva: IASC. www.alnap.org/resource/9975.aspx

ICRC and IFRC (2008) 'Guidelines for assessment in emergencies'. Geneva: ICRC/IFRC. www.alnap.org/resource/6404.aspx

IFRC (2011) 'Project/programme monitoring and evaluation (M&E) guide'. Geneva: IFRC. www.alnap.org/resource/8542.aspx

IFRC (2013) 'World disasters report.' Geneva: IFRC. www.alnap.org/resource/10054.aspx

Jacobsen, K. and Furst Nichols, S. (2011) 'Developing a Profiling Methodology for Displaced People in Urban Areas'. Boston: Feinstein International Center and Tufts University. www.alnap.org/resource/6306.aspx

Jaspars, S. (2006) 'From Food Crisis to Fair Trade: Livelihoods analysis, protection and support in Emergencies'. Series no 3. Oxford: Oxfam. www.alnap.org/resource/9842.aspx

Johnson et al. (2009) 'Research on evaluation use: a review of the empirical literature from 1986 to 2005' in *American Journal of Evaluation* 30(3): 377–410. www.alnap.org/resource/9977.aspx

Jones, H. (2012) 'Promoting evidence-based decision-making in development agencies'. London: ODI. www.alnap.org/resource/9978.aspx

Jones, H. and Mendizabal, E. (2010) 'Strengthening learning from research and evaluation: going with the grain'. London: ODI. www.alnap.org/resource/8429.aspx

Kitson, A. (2002) 'Recognising relationships: reflections on evidence-based practice' in *Nursing Inquiry* 9(3): 179–86. www.alnap.org/resource/9979.aspx

Knox Clarke, P. (2013a) 'Measuring the effectiveness of humanitarian action'. Presentation at *OCHA roundtable on Humanitarian Effectiveness*, March 2013. www.alnap.org/resource/9980.aspx

Knox Clarke, P. (2013b) 'Who's in Charge Here?' A literature review of approaches to leadership in humanitarian operations'. London: ALNAP/ODI. www.alnap.org/resource/8640.aspx

Laybourne, C. and Obrecht, A. (2013) 'Show me the money: generating evidence from and for accountability in humanitarian work', Presented at, *Evidence and Knowledge in Humanitarian Action: ALNAP 28th Annual Meeting*, Washington, 5–7 March. www.alnap.org/resource/8018.aspx

LeCompte, M. D. and Goetz, J. P. (1982) 'Problems of Reliability and Validity in Ethnographic Research'. *Review of Educational Research* 52(1): 31–60. www.alnap.org/resource/9981.aspx

Leeuw, Frans L. (2012) 'Theory Based Evaluation'. www.alnap.org/resource/9982.aspx

Levine et al. (eds) (2011) 'System failure ? Revisiting the problems of timely response to crises in the Horn of Africa: HPN Network Paper'. London: HPN/ODI. www.alnap.org/resource/6260.aspx

Macrae, J. (2013) 'Towards an evidence agenda in humanitarian action: reflections on the role of donors in investing in research and innovation in the sector', Presented at, *Evidence in Humanitarian Response: ALNAP 28th Annual Meeting*, Washington, 5–7 March. www.alnap.org/resource/9843.aspx

Mahmood et al. (2010) 'Guidelines on Data Issues in Humanitarian Crisis Situations'. New York: UNFPA. www.alnap.org/resource/9983.aspx

Majid, N. (2011) 'Early Warning – Delayed Response?'. Geneva: International Federation of the Red Cross. www.alnap.org/resource/6275.aspx

Maxwell, D. and Watkins, B. (2003) 'Humanitarian Information Systems and Emergencies in the Greater Horn of Africa: Logical Components and Logical Linkages' in *Disasters* 27(1): 72–90. www.alnap.org/resource/9984.aspx

Maxwell et al. (2012) 'What Drives Program Choice in Food Security Crises? Examining the Response Analysis Question: Special Edition on Impacts of Innovative Food Assistance Instruments. Boston: Feinstein international. www.alnap.org/resource/9985.aspx

Maxwell et al. (2013) 'HPN Network Paper: Response analysis and response choice in food security crises: a roadmap', 44(73). London: HPN/ODI. www.alnap.org/resource/9986.aspx

Mays et al. (2005) 'Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field' in *Journal of Health Services Research & Policy* 10 (Suppl 1): 6–20. www.alnap.org/resource/9987.aspx

Mazurana et al. (2011) 'Sex and age matter: improving humanitarian response in emergencies'. Boston: Feinstein International Center, Tufts University. www.alnap.org/resource/8144.aspx

Mills, E. J. (2005) 'Sharing evidence on humanitarian relief' in *British Medical Journal* 331: 1485–86. www.alnap.org/resource/9988.aspx

Minear, L. (2002) *The Humanitarian Enterprise: Dilemmas and Discoveries.* Bloomfield: Kumarian Press. www.alnap.org/resource/9994.aspx

Mitchell, G. J. (1999) 'Evidence-based practice: critique and alternative view' in *Nursing Science Quarterly* 12(1): 30–35. www.alnap.org/resource/10000.aspx

Morra Imas, L. and Rist, R. (2009) 'The Road to Results: designing and conducting effective development evaluations'. Washington: The International Bank for Reconstruction and Development and World Bank. www.alnap.org/resource/8470.aspx

MSF Vienna Evaluation Unit (2012) 'Evaluation Manual: A handbook for Initiating, Managing and Conducting Evaluations in MSF'. Vienna : MSF. www.alnap.org/resource/10004.aspx

Nicholson, N. and Desta, S. (2010) 'Evaluation of the Enhanced Livelihoods in Mandera Triangle and Southern Ethiopia 2007–2009'. Washington: USAID and ELMT/ELSE. www.alnap.org/resource/10009.aspx

OCHA (2007) 'Management response matrix to the Intermediate Review of the Central Emergency Response Fund'. New York: OCHA/CERF. www.alnap.org/resource/10012.aspx

OCHA (2012) 'Evaluation Initiale Rapide Multi-Cluster sur les inondations au Moyen Chari, Tandjilé, Mayo Kebbi Est – Tchad'. N'Djema: OCHA. www.alnap.org/resource/10013.aspx

OCHA (2013) 'Meeting Summary: Consultative Workshop Humanitarian Effectiveness'. New York: OCHA. www.alnap.org/resource/10014.aspx

OECD-DAC (2002) 'Glossary of key terms in evaluation and results based management'. Paris: OECD/DAC. www.alnap.org/resource/8489.aspx

Oxfam (n.d.-a) 'How are effectiveness reviews carried out?'. Oxford: Oxfam. www.alnap.org/resource/8473.aspx

Oxfam (n.d.-b) 'Monitoring, evaluation, accountability and learning'. Oxford: Oxfam. www.alnap.org/resource/10015.aspx

Oxfam (n.d.-c) 'Oxfam GB evaluation guidelines'. Oxford: Oxfam. www.alnap.org/resource/10016.aspx

Oxfam (n.d.-d) 'Rough guide to monitoring and evaluation in Oxfam GB'. Oxford: Oxfam. www.alnap.org/resource/10018.aspx

Oxfam (2004) 'Evaluation of Oxfam GB's Food Aid and Food Security Emergency Intervention in Mauritania'. Oxford: Oxfam and Acacia Consultants Ltd. www.alnap.org/resource/10017.aspx

Pantuliano, S. and Wekesa, M. (2008) 'Improving Drought Response in Pastoral Areas of Ethiopia, Somali and Afar Regions and Borena Zone of Oromiya Region'. London: HPG/ODI. www.alnap.org/resource/7510.aspx

Parker, J. and Sanderson, D. (2013) 'Reviewing the quality of evidence in humanitarian evaluations', Presented at, *Evidence and Knowledge in Humanitarian Action: ALNAP 28th Annual Meeting*, Washington, 5–7 March. www.alnap.org/resource/8019.aspx

Peppiatt et al. (2000) *Buying Power: The Use of Cash Transfers in Emergencies.* London: The British Red Cross. www.alnap.org/resource/10019.aspx

Piccioto, R. (2012) 'Probing the paradox of the RCT craze in international development', *NGO Performance,* 24 May. www.alnap.org/resource/10020.aspx

Poole, L. and Primrose, J. (2010) 'Southern Sudan: Funding according to need'. Wells: Development Initiatives. www.alnap.org/resource/10021.aspx

Proudlock et al. (2009) 'Improving humanitarian impact assessment: Bridging theory and practice' in ALNAP Review of humanitarian action in 2009. London: ALNAP/ODI. www.alnap.org/resource/5663.aspx

Puett, C. and Salpeteur, C. (2013) 'Cost-effectiveness of preventing child morbidity with ready-to-use food in urban Chad'. Presented at, *Evidence and Knowledge in Humanitarian Response: ALNAP 28th Annual Meeting*, Washington, 5–7 March. www.alnap.org/resource/8009.aspx

Ramalingam et al. (2009a) 'Innovations in international humanitarian action' in ALNAP Review of humanitarian action in 2009. London: ALNAP/ODI. www.alnap.org/resource/5664.aspx

Ramalingam et al. (2009b) 'Counting what counts : performance and effectiveness in the humanitarian sector'. in ALNAP Review of humanitarian action in 2009. www.alnap.org/resource/5666.aspx

Ravallion, M. (2009) 'Evaluation in the practice of development,' in *The World Bank Research Observer* 24(1): 29–53. www.alnap.org/resource/10022.aspx

Ravallion, M. (2011) 'Knowledgeable bankers? The demand for research in World Bank operations'. World Bank Research Working Paper No.5892. Washington: World Bank. www.alnap.org/resource/10023.aspx

Redmond et al. (2010) 'A Qualitative and Quantitative Study of the Surgical and Rehabilitation Response to the Earthquake in Haiti' in *Prehospital and Disaster Medicine* 26(06): 449–456 www.alnap.org/resource/7446.aspx

Robson et al. (2001) 'Guide to evaluating the effectiveness of strategies for preventing work injuries: how to show whether a safety intervention really works'. Cincinnati: CDC and NIOSH. www.alnap.org/resource/10024.aspx

Rogers, P. J. (2009) 'Learning from the Evidence about Evidence-Based Policy. Strengthening Evidence-Based Policy in the Australian Federation', in *Vol. 1: Roundtable Proceedings*. Canberra: Australian Government Productivity Commission. www.alnap.org/resource/10025.aspx

Sackett et al. (1996) 'Evidence based Medicine: What it is and What it isn't' in *British Medical Journal* 312(5): 71–2. www.alnap.org/resource/10026.aspx

Sadler, K. and Catley, A. (eds). (2009) 'Milk Matters: A literature review of pastoralist nutrition and programming responses'. Boston: Feinstein International Center, Tufts University and Save the Children. www.alnap.org/resource/10027.aspx

Sanderson et al. (2012) Responding to urban disasters'. ALNAP Lessons Paper. London: ALNAP/ODI. www.alnap.org/resource/7772.aspx

Sandison, P. (2006) 'The utilisation of evaluations' in *ALNAP Review of humanitarian action 2005*. London: ALNAP/ODI. www.alnap.org/resource/5225.aspx

Schwandt, T. (2009) 'Towards a practical theory of evidence for evaluation', in Donaldson et al. (2009) *What counts as credible evidence in applied research and evaluation practice*. Thousand Oaks: Sage. www.alnap.org/resource/10028.aspx

Scriven, M. (2009) 'Demythologising Causation and Evidence', in Donaldson et al. (2009) *What counts as credible evidence in applied research and evaluation practice*. Thousand Oaks: Sage. www.alnap.org/resource/10028.aspx

Seaman et al. (2000) 'The Household Economy Approach: a resource manual for practitioners'. London: Save the Children. www.alnap.org/resource/10029.aspx

Segone, M. (2009) 'Enhancing evidence-based policy-making through country-led monitoring and evaluation systems' in Segone, M. (ed.) *Country-led monitoring and evaluation systems*. Geneva: UNICEF. www.alnap.org/resource/8567.aspx

Shaxson, L. (2005) 'Is your evidence robust enough? Questions for policy makers and practitioners' in *Evidence & Policy: A Journal of Research, Debate and Practice* 1(1): 101–12. www.alnap.org/resource/10030.aspx

Shaxson, L. (2012) 'Expanding our understanding of K*(KT, KE, KTT, KMb, KB, KM, etc.)'. A concept paper emerging from the *K* (Kstar) Initiative conference,* 24–27 April, Hamilton, Canada. www.alnap.org/resource/10031.aspx

Silverman, W.A. and Sackett, D. L. (1999) *Where's the evidence? Debates in modern science*. Oxford: Oxford University Press. www.alnap.org/resource/10032.aspx

Slim, H. (2012) 'IASC Real-Time Evaluation of the Humanitarian Response to the Horn of Africa Drought Crisis Synthesis Report'. New York: IASC. www.alnap.org/resource/7761.aspx

Smith et al. (2011) 'New Technologies in Cash Transfer Programming and Humanitarian Assistance' Oxford: CaLP. www.alnap.org/resource/9634.aspx

Spence et al. (2003) 'Quality in qualitative evaluation: a framework for assessing research evidence'. London: Government Chief Social Researcher's Office. www.alnap.org/resource/10033.aspx

Steen et al. (2013) 'Synthesis of Mixed Method Impact Evaluations of the Contribution of Food Assistance to Durable Solutions in Protracted Refugee Situations'. Rome: WFP. www.alnap.org/resource/7997.aspx

Stern et al. (2012) 'Broadening the range of designs and methods for evaluationst'. DFID Working Paper 38. London: DFID. www.alnap.org/resource/8196.aspx

Tache, B. (2010) 'Participatory Impacts Assessment of Drought Reserve Areas in Guji and Borana Zones', Oromia Region: Save the Children USA. www.alnap.org/resource/9940.aspx

Telford, J. and J. Cosgrave (2006) 'Joint evaluation of the international response to the Indian Ocean tsunami: Synthesis report'. ALNAP and TEC/ODI. www.alnap.org/resource/3535.aspx

Tessitore, S. (2013) 'Like a good trip to town without selling your animals: A study of FAO Somalia's Cash for Work programme'. Rome: FAO. www.alnap.org/resource/8821.aspx

UNHCR (2010) 'UNHCR's evaluation policy'. Geneva: UNHCR. www.alnap.org/resource/10034.aspx

UNICEF (2004) 'UNICEF evaluation report standards'. New York: UNICEF. www.alnap.org/resource/10035.aspx

UNICEF (2007) 'Programme policy and procedure manual: programme operations'. New York: UNICEF. www.alnap.org/resource/10036.aspx

UNOCHA and Pakistan NDMA (2012) 'Multi-sector Initial Rapid Assessment (MIRA) Report Pakistan Floods 2012'. Islamabad: Assessment Working Group Pakistan. www.alnap.org/resource/10038.aspx

Valid International (2012) 'IASC Real Time Evaluation of the response to the Horn of Africa Drought: Somalia'. Oxford: Valid International. www.alnap.org/resource/7506.aspx

Van de Putte, B. (2000) 'The Utilisation of Evaluation Recommendations in Medecins Sans Frontieres – Holland: A Study of 10 Evaluation Reports (1997–1999)'. Amsterdam: MSF-H. www.alnap.org/resource/10040.aspx

Van de Putte, B. (2001) 'Follow-up to Evaluations of Humanitarian Programmes'. London: ALNAP/ODI. www.alnap.org/resource/5011.aspx

Venton et al. (2012) 'The Economics of Early Response and Disaster Resilience: Lessons from Kenya and Ethiopia: Economics of Resilience Final Report'. London: DFID and UKaid. www.alnap.org/resource/7758.aspx

VSF (2009) 'Meat and Milk Voucher project (IMPACT I and II)', Clarke and Fison. Bahr al Gazal, South Sudan: VFS. www.alnap.org/resource/10042.aspx

Walden, V. (2013) 'Improving quality of humanitarian programmes through the use of a scoring system: the Humanitarian Indicator Tool', Presented at, *Evidence and Knowledge in Humanitarian Action: ALNAP 28th Annual Meeting*, Washington, 5–7 March. Oxford: Oxfam. www.alnap.org/resource/8024.aspx

Walker, P. (2013) 'Cracks in the machine: is the humanitarian system fit for purpose?', Presented at, *Evidence and Knowledge in Humanitarian Action: ALNAP 28th Annual Meeting*, Washington, 5–7 March. www.alnap.org/resource/8014.aspx

Willet, B. (2013) 'Mission metrics: An agency wide measurement approach for understanding emergency response', Presented at *Evidence and Knowledge in Humanitarian Action: ALNAP 28th Annual Meeting*, Washington, 5–7 March.www.alnap.org/resource/8025.aspx

WFP (n.d.) 'Monitoring and Evaluation Guidelines. Module: How to design a Results-Oriented M&E Strategy for EMOPs and PRROs'. Rome: WFP. www.alnap.org/resource/10045.aspx

WFP (2005) 'Summary report on WFP follow-up to recommendations'. Rome: WFP. www.alnap.org/resource/10048.aspx

World Vision Pakistan (2011) 'End of Programme Evaluation Report for DEC-funded WV Relief Program in Sindh (July 2011 – Phase 1)'. Islamabad: Sustainable Solutions Ltd. www.alnap.org/resource/6148.aspx

Young, J. and Court, J. (2004) 'Bridging Research and Policy in International Development: An Analytical and Practical Framework'. London: ODI. www.alnap.org/resource/8430.aspx

Zhang et al. (2002) 'A knowledge management framework for the support of decision making in humanitarian assistance/disaster relief' in *Knowledge and Information Systems* 4(3): 370–85. www.alnap.org/resource/10052.aspx

ALNAP
Overseas Development Institute
203 Blackfriars Road
London SE1 8NJ
United Kingdom

T + 44 (0)20 7922 0388
F + 44 (0)20 7922 0399
E alnap@alnap.org

ALNAP would like to acknowledge the financial support of USAID in carrying out this initiative.