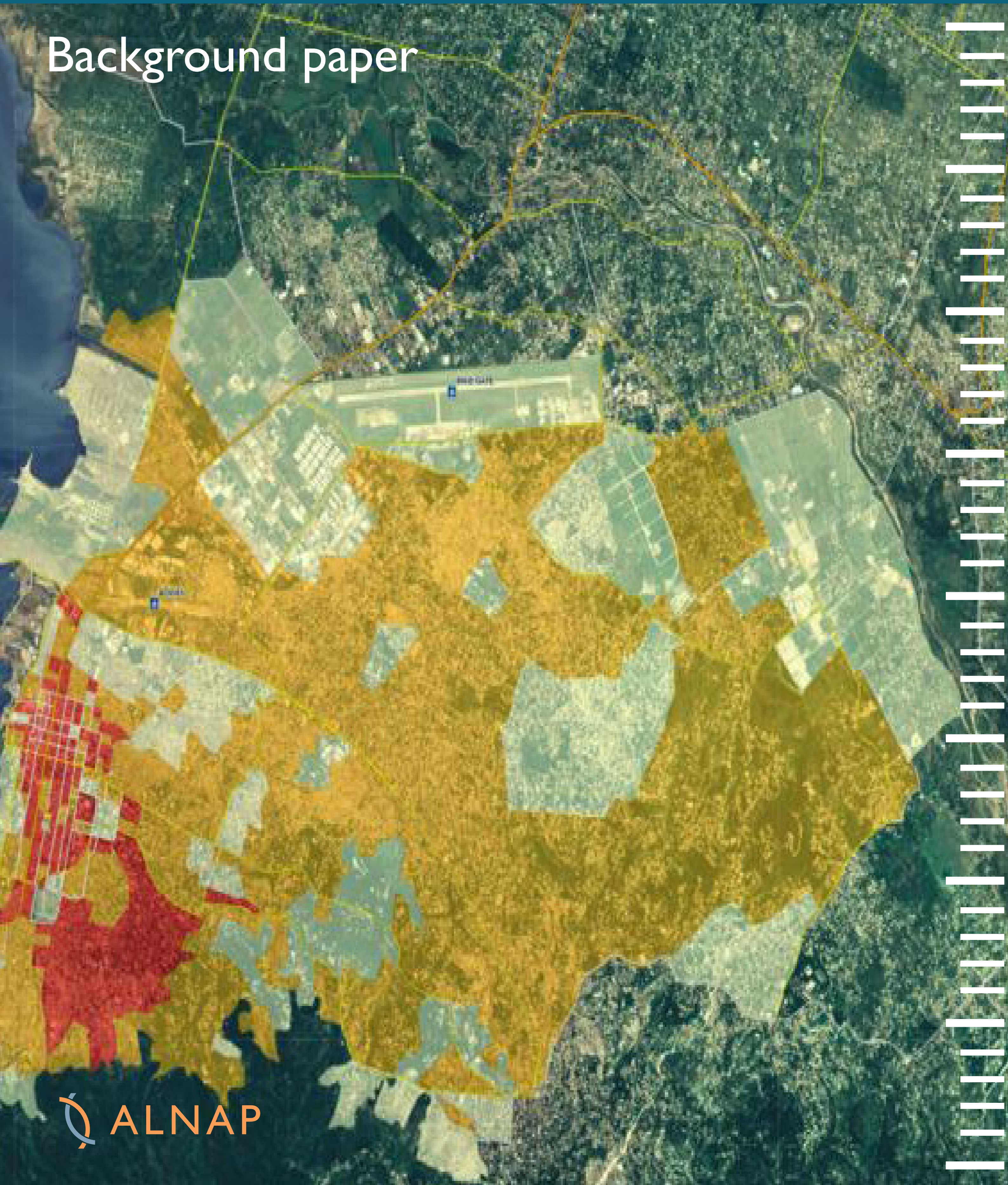
28th ALNAP MEETING

WASHINGTON, D.C.
5-7 MARCH 2013

# EVIDENCE & KNOWLEDGE

## IN HUMANITARIAN ACTION

Background paper



ALNAP

## Acknowledgements

This paper was commissioned by the ALNAP Secretariat as a background paper for the 28th Annual ALNAP meeting on Evidence and Knowledge in Humanitarian Action. It was written by James Darcy and Paul Knox Clarke, with research assistance from Miriam Bradley and Alexandra Warner. The authors would like to express their thanks to Mary Dixon-Woods, Jeff Crisp, Andrew Catley, John Mitchell, Haley Bowcock, Lili Mohidin, Sharon Truelove, Merry Fitzpatrick, Paul Harvey, Yoann Tuzzolino and Francesca Bonino, all of whom provided helpful input. The authors remain fully responsible for the views expressed and for any errors or omissions in the text.

## CONTENTS

# SECTION 1 – BACKGROUND

## 1.1 Why this topic?

The subject of this background paper – and of the 28th ALNAP Annual Meeting – is how evidence and knowledge inform policy and practice in the humanitarian sector. This is not a new topic by any means, but it raises questions that have become increasingly pressing for the sector in the past few years. The mid-1990s saw a period of NGO-led self-reflection and standard-setting, and a subsequent UN-led focus on better coordination and leadership. This reflected general concerns with the overall performance of the international humanitarian system and the need for greater accountability. One strand of that concern and of parallel donor-led initiatives like the Good Humanitarian Donorship agenda, has been the quality of the data and analysis that underpins crisis responses, the extent to which those responses are genuinely 'needs based', and whether their effectiveness can be demonstrated through evidence. Increasingly this has put a spotlight on the diagnostic and predictive analysis generated by early warning, needs assessment and monitoring processes. It has also placed a focus on the more retrospective judgements of evaluation processes, particularly as they concern impact and effectiveness. More generally, it raises questions about current humanitarian policy and practice and the quality of evidence on which they are based.

**1.2** Just as important as the availability and quality of evidence is the question of how – or indeed whether – such evidence is ***used*** by decision-makers. Recurrent collective failures to respond decisively in the face of strong evidence of impending crisis (notably from famine early warning systems in sub-Saharan Africa) highlight the point that ***generating*** such evidence is only one part of the challenge. This is true also of evidence from past experience: a recurrent theme of evaluations is that the international system and individual organisations struggle to learn lessons and apply evidence from past experience to current practice (Sandison, 2006; Hallam, 2011).  The way in which evidence and knowledge is communicated, assimilated and acted upon by decision-makers is central to this.

**1.3** It would be misleading to suggest that no progress has been made over the past two decades in relation to the generation and use of evidence and knowledge in the humanitarian sector, whether through organisational learning processes or through more collective endeavours of research, assessment, codification and standard-setting (Walker and Purdin 2004; Young and Harvey 2004). In this respect the sector as a whole certainly looks more professional than it did 15 years ago (Barnett 2005). For example there has been the application of inter-organisational minimum standards like Sphere, and work on joint assessment and analysis within and between sectoral clusters. But in most areas of 'diagnostic' and 'learning' practice the humanitarian sector appears weak compared to other sectors, including the wider development sector. This cannot be entirely explained by the peculiar nature of the humanitarian enterprise and the constraints of working in crisis contexts. Underlying this paper and the ALNAP meeting is the sense that much humanitarian practice and policy has developed with only limited reference to the evidence base. As a result we may not be working as effectively as we could. Many feel that the sector can and should do better, not least because humanitarians owe it to those they seek to assist to deal in actual – rather than hypothetical – problems and outcomes.

**1.4** Various recent policy developments make these issues particularly pressing at present. Some of these concern donor expectations about the demonstration of results and of 'value for money'. The humanitarian sector is increasingly subject to the same pressures as other areas of public spending in this regard. The expectation in the medical and public health spheres, and in public policy more generally, is that practice should be justified against established 'best practice' and that neither existing policy nor the authority of experts should be immune to challenge on such grounds. What constitutes best practice, and the methods by which best practice can be best identified, is a matter of debate in the wider social sector. Humanitarians are felt by many to have lagged in this debate, even compared to their development colleagues. We may have something to learn from the ways in which other sectors have responded to these pressures. In this paper, we consider some of the more relevant points of comparison with other sectors.

**1.5** This paper aims to help structure a dialogue about these issues. It consists in part of a 'stock take' of current practice in the humanitarian sector with regard to the generation and use of evidence, highlighting apparent strengths and weaknesses of current practice. It touches on some of the more relevant aspects of current practice in other sectors, including medicine, public health and law. It raises questions about incentives and disincentives for the use of evidence in the humanitarian sector. It also considers some of the ways in which evidence-informed practice might be strengthened, without attempting to provide more than indicative answers to these questions.

**1.6** The paper is concerned with evidence and knowledge as seen from various perspectives: that of the person affected by crisis; that of the humanitarian practitioner concerned with response decisions, or with the design, implementation and monitoring of specific programmes; that of the evaluator or researcher, concerned with testing particular programmes or strategies and considering what generic lessons can be learned; and that of the policy-maker or manager concerned with devising strategy, policy and standards. These may of course be overlapping concerns. The point is that we are concerned ***both*** with evidence that is context-specific and necessary to inform real-time response decisions; ***and*** with evidence that supports more general conclusions about (for example) the relative merits of different programme approaches to different types of crisis. In short, we are concerned with evidence and knowledge in relation both to practice and to policy.

## Definitions

For the purposes of this paper, we adopt the following working definitions of key terms:

**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 such as to inform a true understanding of a particular topic.
**Information:** any data that may inform understanding or belief, presented in a context that gives it meaning. Information may be true or false.
**Evidence:** information that helps substantiate or prove the truth of a proposition.
Proof literally means 'showing the merit of' a proposition. Giving proof involves providing sufficient evidence to demonstrate the truth of a given proposition. The probative value of evidence relates to the extent to which it goes towards proving a proposition.
**Hypothesis:** a proposed explanation for a phenomenon. More generally used as equivalent to an argument or theory.
**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.
**Accuracy:** how close a measurement of a quantity is to that quantity's actual (true) value.
**Precision:** (of a measurement system) the degree to which repeated measurements under unchanged conditions show the same results.
**Inference:** the process of reaching conclusions as an extension of known facts or stated premises. Inference is involved (for example) in drawing conclusions about a population from a sample. Causal attribution is based on inference.
**Bias:** any form of systematic (non-random) error.
**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:** used of analysis and hypotheses ('based on sound argument and inferences') and experimental design (able to demonstrate what it claims). Often the concern is with the validity of attribution of an effect to a particular cause or intervention. In scientific studies, the terms internal and external validity refer to the extent to which different types of causal inference are warranted. Internal validity concerns the validity of causal conclusions within a particular study or situation. External validity concerns the extent to which causal relationships identified in a particular situation are generalisable to other situations.

# SECTION 2 - THEORY AND CONCEPTS

## 2.1 Understanding terms, concepts and theories

**2.1.1** This paper is based on a few core concepts that it is 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*. For our purposes, we take it to mean a 'justified true belief'– in other words, a belief that is in accordance with observable facts, where there are grounds for believing it to be so. This definition leaves plenty of room for debate about the nature and truth of the belief in question, who actually believes it and on what basis. In practice, what constitutes 'knowledge' in a particular field may depend on the prevailing consensus view at a given time, i.e. that which is generally *agreed* to be true. In the context of this paper, knowledge is understood to derive either from direct observation or from an accumulation of evidence such as to inform a true understanding of a particular topic. 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. The data in question or the associated belief may be true or false.

***Evidence*** is information of a particular kind. We understand it here to mean true or credible information (quantitative or qualitative) that helps demonstrate the truth or falsehood of a given proposition. So for example, nutritional data gathered using a valid method of nutritional assessment may constitute compelling evidence for the existence of a nutritional crisis in a given context. An anecdotal report of large numbers of wasted children in a given region is information that *may* also constitute evidence to support the same conclusion, though of a less compelling kind.[1] The conclusions from a rigorously conducted series of trials may provide a strong evidence base for policy formulation; the evidence from a single programme evaluation much less so. Thus there may be degrees of ***strength*** of evidence, a function mainly of the *accuracy* of the information in question (or the reasons for thinking it true) and its *significance*: what it actually tells us about a situation or approach.[2]

The *threshold* for evidence – what evidence we need before we are prepared to accept and act on a given proposition – may vary from context to context. 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.

**2.1.2** Within the field of epistemology (concerning theories of knowledge), 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 Western scientific thought for the past 400 years. As Donald Campbell puts it:[3] 'Science requires a disputatious community of "truth seekers". The norms of science are explicitly anti-authoritarian, anti-traditional, anti-revelational and pro-individualistic … Old beliefs are to be systematically doubted until they have been reconfirmed by the methods of the new science'.

> *Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted.*
>
> Albert Einstein

The empirical tradition in Europe has philosophical and scientific roots – from radical thinkers and experimental scientists ('natural philosophers') like Francis Bacon and Galileo in the 17th century to the scientists of the later Enlightenment period. These scientists and their philosophical contemporaries like Locke, Spinoza, Berkley and Hume, all privileged experience-based reasoning over authority and tradition. In the 19th century, Comte and Durkheim recast empiricism as *positivism* in applying it to the emerging social sciences. For the logical positivists of the 20th century, propositions that could not be tested using observation and experience, unless they were true by definition or logical inference, were simply meaningless. That included all metaphysical propositions.

[1] Less compelling both because it may not be true (e.g. mistaken observation) and because it may not be possible to substantiate a more general proposition about malnutrition from such an observation.

[2] In place of 'significance' we sometimes use the more specific term 'probative value' below, a term borrowed from the legal sphere. The criteria by which the strength of evidence is judged are considered in detail later in this section.

[3] Cited in Gerring 2012, p.9

**2.1.3** Social scientific approaches are not 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, and with it the idea of strict objectivity; and 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.[4]

In spite of these challenges and qualifications to empiricism and positivism, the Western scientific method – involving the formulation and testing of hypotheses through experimentation and observation – continues to be a dominant influence on thinking about knowledge and evidence in the social as well as the natural sciences. This involves the search for true (and sometimes generalisable) propositions on the basis of which deductions and predictions can be confidently made, and on which policy and practice can be safely founded. Sometimes the truth is counter-intuitive or runs counter to accepted wisdom (or vested interests), with the result that it struggles to gain acceptance.

### The early use of clinical trials

Practitioners like James Lind in the 18th century showed the value of carefully conducted and well documented trials in challenging received wisdom. Lind, an English ship's doctor and an early advocate of preventive medicine, pioneered the use of controlled clinical trials to prove (*inter alia*) that citrus fruits could be used to cure scurvy.

Two months into a sea voyage when the ship was afflicted with scurvy, Lind divided 12 scorbutic (scurvy-affected) sailors into six groups of two. They all received the same diet but, in addition, group one was given a quart of cider daily, group two twenty-five drops of elixir of vitriol (sulphuric acid), group three six spoonfuls of vinegar, group four half a pint of seawater, group five received two oranges and one lemon, and the last group a spicy paste plus a drink of barley water. The treatment of group five stopped after six days when they ran out of fruit, but by that time one sailor was fit for duty while the other had almost recovered. Apart from that, only group one also showed some effect of its treatment.

The medical establishment ashore continued to be wedded to the idea that scurvy was a disease of putrefaction. It could not account for the benefits of citrus fruits and dismissed the evidence in their favour as unproven and anecdotal. In the Navy however, experience had convinced many officers and surgeons that citrus juices provided the answer to scurvy even if the reason was unknown. Lind had no theory to explain his results, but subsequent trials led to the same outcomes – and eventually forced a change of policy. Lind's approach, with modifications like randomisation and 'blinding', still forms the basis of good medical research practice today.

[4] For an interesting perspective on how these issues shape current debates about the use of evidence in the development sector, see the recent blog discussion hosted by Duncan Green of Oxfam at: http://www.oxfamblogs.org/fp2p/?p=13344

**2.1.4** While formulation and testing of hypotheses is an application of the Western scientific method that pervades the social sciences, the nature of the hypotheses and the kinds of evidence taken to demonstrate them are inevitably complicated by behavioural and context-specific factors. This is relevant to the application of such methods to humanitarian contexts. Crisis contexts do not present laboratory conditions in which variables can be controlled and where individuals and groups always behave in the same way. The possibility of using experimental approaches – and particularly controlled trials – to generate evidence in the humanitarian context is explored further in Section 3. Here we simply note that the application of standard scientific approaches, even those used in other areas of social policy, is not straightforward.

**2.1.5** The modern humanitarian enterprise is characterised by the co-existence of multiple disciplines: medicine, public health, economics, engineering, agriculture and anthropology, to name just a few. Each has its own terminology, governing concepts and theoretical frameworks. Some are more traditionally 'scientific' or statistically based than others. Each tends to privilege certain kinds of information and evidence over others, with a greater or lesser emphasis on qualitative or quantitative methods and indicators. Yet of course, people's lives do not divide neatly into these categories. Given the interaction in practice of the various factors involved in people's lives, mixed-method approaches to evidence gathering and analysis are generally required to make sense of them. Unless we consider behavioural and perceptual issues alongside technical ones, we are likely to go badly wrong in our understanding of crisis situations and of the appropriate responses to them.

This then raises some tough methodological questions: what is the framework of analysis that allows us to bring together evidence of quite different kinds from different disciplines? Do we have enough commonality of language and concepts to allow this? Or is it simply a matter of presenting decision-makers with different streams of evidence and relying on them to use their own judgement in deciding how to weigh one kind of factor against another? While considerable progress has been made in agreeing common methods, standards and indicators within each discipline in the humanitarian context, there is little in the way of unifying analytical frameworks across disciplines.[5] We rely heavily on the synthesis of evidence made by those presenting the case for action.

**2.1.6** Over the past decade or more, the demand for public policy to be more firmly grounded in evidence has grown. 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 (Davies et al. 2000; Nutley, Walter and Davies 2007; NAO 2003). The movement arose out of twin concerns. First that systematic knowledge and analysis was often not utilized by policy-makers; and second that policy makers had no way of judging the trustworthiness of the findings with which they were bombarded by academics, pressure groups and lobbyists.' (ibid, para 2.17). The influence of the EBP movement is now being felt in the humanitarian sector, with an increased desire to use scientific methods to test 'orthodox' approaches. However, establishing experimental conditions in humanitarian contexts can often be hard to do, and this scientific approach raises concerns from some quarters that these methods can overlook hard to measure, but essential, elements of cultural and political reality.

Sections 3 and 4 below explore the ways in which evidence is generated and used both by humanitarian practitioners and by policy-makers.

[5] Perhaps the greatest progress in this regard has been made across the fields of public health, nutrition and food security/livelihoods, by UNICEF and others. Taking mortality, morbidity and malnutrition as the basic outcomes of concern but allowing for the effects of social, economic and political factors – including access to services – a reasonably coherent explanatory model can be built.

## 2.2 Evidence for what? Testing humanitarian propositions

**2.2.1** As described above, 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 so closely linked to the idea of propositions, it is important 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 three linked types of proposition underlying humanitarian interventions by international agencies. Essentially, these are as follows:

A. Propositions about the existence of an actual or potential crisis;
B. Propositions about 'what works' in preventing or mitigating crises of this kind;
C. Propositions about the most appropriate response to a particular crisis.

Each requires evidence to support it, but the kinds of evidence involved for each may be quite different. We consider these propositions and their evidential requirements in more detail below.

**2.2.2** The three proposition types above need to be stated more exactly before they can be properly analysed. Here we present them in expanded form.

### Proposition type A (**Diagnostic** or problem statement)

(i) **A situation exists that is 'critical' for those affected**, as gauged against agreed indicators of crisis.
(ii) **The situation will lead to (continued) catastrophic human outcomes without humanitarian intervention.**[6]

This might be called the trigger proposition. It depends on demonstrating that the qualifying conditions for a 'crisis' situation have been met, e.g. a nutritional crisis based on raised (or rising) observed levels of acute malnutrition. This type of proposition is generally based on a description of *symptoms* such as disease morbidity, food consumption or the number of houses destroyed. In many cases, these symptoms are compared with accepted crisis thresholds and indicators. 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 (e.g. epidemics relating to lack of water, sanitation, hygiene), food security and shelter. Many of these sectors have their own indicators and thresholds, although in some cases interventions 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 fact that in some cases thresholds do not exist, raises questions about what constitutes a crisis and what evidence is needed to demonstrate it.

Type A propositions – generally made on the basis of evidence from needs assessment or early warning processes – may go beyond description of symptoms to identify the proximate causes and the expected *development* of the crisis. Indeed, a full 'diagnosis' (or prognosis) requires such elements, though in practice they are often either implicit or absent from the problem statement.

[6] This can alternatively be stated as a conditional: 'If there is no intervention, then the result will be catastrophic'.

### Proposition type B **(Effectiveness of response)**

**Intervention of a specified type will be effective in preventing or mitigating the effects of the crisis** (or any such crisis) in defined ways.[7]

Propositions of this type are at the core of response choice: deciding how to intervene in a specific situation. They are also important in evaluation, when considering whether a response was effective. They require evidence about ***what works / worked, or what will work in the context***, to change outcomes without causing undue harm.

Type B propositions are not usually ***absolute*** in their description of outcomes; rather, they tend to describe outcomes ***relative*** to the results of non-intervention. They also tend to be based on probabilities. In effect, we are generally saying that if we take an action (say the introduction of targeted cash transfers) then a certain number of people are less likely to become food-insecure than if we do nothing.

In practice there is generally a third, business-oriented proposition linked to type B propositions, concerning the choice of response and responder. Typically this has two main parts:

### Proposition type C **(Appropriateness of response)**

**(i)** **The proposed intervention is the most appropriate available in the context**, taking into account likely effectiveness, local preferences, alternative options, cost, timeliness, etc. ('Best option')

**(ii)** **The intervention can be delivered** on the basis proposed, meeting agreed minimum standards. ('Feasibility')

While Type B propositions compare the effectiveness of an intervention with doing nothing , Type C, or appropriateness propositions, depend on demonstrating why the proposed intervention is preferred among the options for response available. In the past, many agencies may have used a fairly limited range of 'default responses', but these are increasingly being challenged by advances in learning about alternative options. In many areas agencies are now required to demonstrate how their performance rates against a broad range of possible interventions, increasing the evidential burden. So 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 beginning to put greater pressure on agencies to justify their choice of response option (Maxwell et al. 2012). The evidence required to substantiate these propositions depends on the criteria used to determine which intervention is 'best': it may, for example, be (cost-)efficiency; (cost-)effectiveness; acceptance by the population; or some combination of these. Each will require different information sets, and different methods for collection and analysis.

**2.2.3** Each of the propositions above may be either ***prospective*** or ***retrospective*** – relating either to current/future or past events – and they may have to be demonstrated either ***before*** or ***after*** an intervention has taken place. While early warning, monitoring and assessment processes are concerned with propositions about the present and likely future, most evaluations are concerned with testing the ***retrospective*** validity of the propositions above: 'There ***was*** a critical situation, and the intervention in question helped mitigate its effects. This was the most appropriate intervention in the circumstances,

[7] This can be stated as: 'If we intervene in this way then catastrophic outcomes will be averted'.

and it was delivered in accordance with best practice.' Evaluations are concerned with the effectiveness and appropriateness[8] of the intervention, and with the way that intervention was delivered.

**2.2.4** Each of these types of proposition requires evidence to demonstrate their truth. Our interest here is in the kinds of evidence involved, principally concerning the evolution of crisis situations and the kinds of response that are effective in responding to them. The generation and use of these two forms of evidence are the main concern of this paper.

For both types A and B, evidence may be direct or indirect. ***Proxy indicators*** provide indirect evidence of the proposition in question and are often used on the grounds that they may be simpler to measure and provide reasonably reliable evidence to support the more general proposition. So for example, dietary diversity or the number of meals consumed by household members per day might be used as proxy indicators of food security. Such indicators are rarely conclusive and they can be overused. A key evidential concern in humanitarian action, particularly for diagnostic (type A) propositions, is the selection of these proxy indicators: where the wrong indicators are selected, it is possible to have good 'truthful' information that is not 'significant' as evidence to prove or disprove the diagnosis.

**2.2.5** Some overall comments can be made on the nature of evidence required for the three different types of proposition. For type A (diagnostic) propositions, evidence is likely to be weighted towards that which describes a situation and which is context-specific. Where generic evidence is used, it is generally to establish what the 'crisis threshold' should be (so that, for example, the threshold for famine is the same in all places and situations) and to explain how the event is likely to unfold, in the light of previous comparable events. Arguably, too many generic assumptions about outcomes are made in the immediate aftermath of a crisis and are not sufficiently tested against the contextual reality through subsequent (re-)assessment and monitoring.

Type B (effectiveness) propositions tend to require evidence of causality, to show that the intervention led to (or will lead to) a relatively better situation. When these are prospective – that is, when deciding what intervention to use – they tend to rely on a body of generic evidence on what has worked in similar situations elsewhere. When they are retrospective, they depend mostly on context-specific evidence of the degree to which the intervention 'worked' in this case, showing a linkage between the intervention and the more positive outcome. This often requires evidence to show what would have happened if there had been no intervention (counterfactual evidence) which can be difficult to find.

For type C (appropriateness) propositions, the questions to be answered – and hence the evidence required – are largely context-specific, but inevitably they depend on consideration of established track-records of particular approaches or agencies. Increasingly, agencies are under pressure to demonstrate why a particular response is the right one on grounds of cost and appropriateness to context. They need to make the 'feasibility' case for the intervention and to say why that agency is in a position to conduct it.

## 2.3 Evidence, proof and the testing of hypotheses

**2.3.1** What constitutes sufficient evidence for humanitarian purposes? Depending on the context and the type of decision to be made, different thresholds for strength and weight of evidence are likely to apply, with a lower threshold being applied to situations requiring an urgent decision about a particular response than to (say) the formulation of a policy about responses in general. We consider here how the

[8] In principle, in considering the question of appropriateness, evaluations ought to concern themselves with the validity of the original diagnosis/prognosis on which the response was based. In practice they do not always do so.

idea of thresholds and weight of evidence is used in two other sectors: law and medicine.

**2.3.2** In the legal sector, evidence is used to demonstrate the existence of facts that go to help establish a particular legal case. Different thresholds of evidence apply in different kinds of case. In most codes of criminal law, a high evidential threshold applies; typically, the evidence has to be such as to prove 'beyond reasonable doubt' that the person in question is guilty of a criminal act. In a civil legal case the threshold is generally lower: e.g. that person needs to show that 'on the balance of probabilities' the defendant is liable in law for breach of obligation (e.g. breach of contract, failure of duty of care).

Circumstantial evidence in law is evidence in which an inference is required to connect it to a conclusion of fact, like a fingerprint at the scene of a crime. By contrast, direct evidence supports the truth of an assertion directly – i.e., without need for any additional evidence or the intervening inference. So if a witness in a murder trial claims to have seen the accused killing the victim, this is direct evidence (though its accuracy has to be tested). If on the other hand they claim to have seen the accused leaving the victim's house with blood on their hands, this is indirect (circumstantial) evidence. On its own, circumstantial evidence leaves open the possibility of more than one explanation. Circumstantial evidence can accumulate into a body of ***corroborating*** evidence in support of one particular inference over another.

**2.3.4** In the medical sector, the term 'evidence' is now most commonly associated with ***evidence-based medicine***, in which the evidence in question concerns the efficacy and safety of a proposed course of treatment for a given medical condition. We consider the potential applications of this line of thinking to the humanitarian sector in Section 4. But the use of evidence in medicine is by no means restricted to this area, just as it is not in the sciences more generally. In particular, the process of ***diagnosis*** and ***prognosis*** involves formulating and testing hypotheses about the nature, causes and likely development of an apparent medical condition. Evidence is used both in formulating the hypothesis ('this child has measles'), drawing on knowledge about how different diseases present; and then in testing that hypothesis through established diagnostic procedures.

**2.3.5** How does the above relate to the humanitarian sphere? We take three main ideas from the legal sphere. The first concerns the idea of a 'threshold of evidence'. Though rarely required to demonstrate the proof of a proposition 'beyond reasonable doubt', humanitarian actors do use the idea of a 'balance of probability' concerning the likelihood of outcomes. This is particularly so in risk analysis and scenario planning where the 'most probable' outcome indicated by the evidence may form the primary basis for planning, while contingency planning against outcomes of lower (but significant) probability and potentially disastrous impact. In policy terms, the evidential threshold for decisions is higher, though rarely as high as 'beyond reasonable doubt' – since the existence of 'reasonable doubt' tends to be a characteristic of crisis contexts.

The second idea – that of circumstantial evidence – has parallels with the use of 'proxy' and other indirect indicators of crisis in the humanitarian context; with the need for corroborating evidence in both. Often in crisis contexts, time and access constraints mean that there is a lack of direct evidence, so that proxy indicators (e.g. numbers displaced, extent of damage to property, changes in consumption patterns or other behaviours) are used to estimate the nature and extent of need. The question arises as to whether the proxy indicator in question is an appropriate one, and whether there is over-reliance on such indicators at a stage where more direct evidence is available.

The third idea we take from the legal sphere concerns the 'probative value' of evidence. We use this here as a more precise term for the ***significance*** of evidence in relation to a given case or proposition.

Evidence has probative value to the extent that it helps prove or support a particular case. In other words, it concerns the question: *what does this evidence tell us* with regard to a given proposition.
From the medical and public health spheres we again take three main ideas, which are explored in Sections 3 and 4. The first concerns the process of testing, diagnosis/prognosis, treatment and re-testing, which is analogous to the process of needs assessment (situation and response analysis), response and outcome monitoring in the humanitarian context. The second idea concerns the 'evidence-based' approach to testing treatments for effectiveness and safety. The third idea, related to the second, concerns hierarchies of evidential strength in relation to the testing of treatments.

### Hierarchies of evidence strength: comparison with the medical sphere

The issue of the quality and strength of evidence depends heavily on the source of evidence and its method of acquisition. Bradt (2009) describes how in evidence-based medicine (EBM) evidence is organised according to a hierarchy of evidence strength. The Centre for Evidence-Based Medicine in Oxford has developed such a hierarchy based upon the method of data acquisition, with systematic review of randomised controlled trials (RCTs) at the top, followed by individual RCTs. Expert opinion is ranked in the lowest (fifth) category, just below 'case series and poor-quality cohort and case-control studies' (fourth category).

Does it make sense to apply such a typology to humanitarian evidence? The differences pointed out above should make us cautious. Most evidence in the humanitarian sphere falls into EBM categories 4 and 5 above, i.e. the weakest of the categories; or even outside the scale altogether. So what should we conclude from this? We may judge that this is simply an inevitable consequence of the nature of humanitarian contexts, where it is either not feasible or not ethical to conduct the kinds of trials that form the basis of the stronger evidence categories in EBM. But this has been disputed by some (see section 3). In any case, it is reasonable to ask what the equivalent hierarchy of evidential strength might be in the humanitarian sphere, allowing that some purposes (e.g. general policy formulation) may require stronger evidence than others. There is an accountability dimension to this. The related performance question here would be: was the best available evidence used to inform the response, or to inform a given policy?

## 2.4 Gauging the quality, strength and significance of evidence

**2.4.1** In trying to substantiate the propositions above, we run into a number of problems. 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? And how do we judge whether the evidence actually supports the proposition in question? Here we consider some of relevant issues concerning the testing of evidence.

### 2.4.2 Testing different types of evidence

Evidence in the contexts we are considering takes two main forms: (i) data, whether quantitative or qualitative, direct or indirect; (ii) the results 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 held up against specific quality criteria.

### 2.4.3 Criteria for assessing the quality of evidence

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 both to describe the quality of data and of the *methods* by which data are collected and analysed. Here we have attempted to summarise the main criteria that might be used for assessing the quality of evidence in humanitarian contexts.

> (i) **Truth** or **accuracy**: whether the evidence corresponds to a real state of affairs. Since truth is often difficult to establish, the credibility of the evidence is often taken as a proxy: credible evidence is that which derives from credible (believable, trustworthy) sources, and for which there are good reasons for believing it to be true. In considering the truth of any information, it is important to consider the biases of those who collect the information, as well as those who provide it.
>
> (ii) **Representativeness** of the evidence: whether (for example) data collected is statistically representative of the wider group; or whether an opinion survey represents the views of a wider group.
>
> (iii) **Significance** or probative value: the significance of the information or analysis in evidential terms, the extent to which it helps demonstrate the truth of a given claim or proposition. This includes its relevance to the claim in question. The presentation of a bald statistic such as '10 per cent prevalence of global acute malnutrition in district X' may tell us little on its own about the existence and nature of a nutritional crisis, unless we also know what the trend is (rising, falling?), what the seasonal norm is for the district in question, related food security and morbidity patterns, etc. Even then, such evidence on its own is likely to tell us little about the appropriate form of response. The issue of significance is of particular concern when we are choosing proxy indicators.
>
> (iv) **Generalisability**: of conclusions (for example, about the physiological consequences of malnutrition, or the acceptability of certain types of shelter for earthquake survivors), whether they can be generalised beyond the context in question to other contexts, and so used as evidence of how a situation will unfold, or of the best type of intervention to use.
>
> (v) **Attribution**: of analysis, whether it demonstrates a clear and unambiguous causal linkage between two conditions or events.

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 which together define the 'robustness' or strength of evidence in policy terms: credibility, reliability, objectivity (lack of bias), rootedness, and generalisability. '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. In other words, are we asking the right questions?

John Gerring, in proposing a unified framework for social scientific methodologies, suggests some criteria that apply to all social science arguments (propositions), including (*inter alia*) 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.'[9] So we should not be surprised that a single list of agreed criteria is hard to produce.

# SECTION 3 – GENERATING KNOWLEDGE AND EVIDENCE IN PRACTICE

## 3.1 Introduction

There are a number of serious challenges to the generation of good evidence in the humanitarian sector. With regard to specific crises, some of these challenges relate to the difficulties of securing ***any*** reliable information from the areas and communities affected in the early days of a rapid-onset disaster, or in highly insecure and volatile environments where access may be limited. When the humanitarian situation is evolving rapidly, information (and so evidence) may quickly become out of date – so that the evidence base has to be renewed regularly and responses adapted accordingly. The data on which evidence is based is generally 'historic', reflecting the situation as it existed days, weeks or even months earlier. What constitutes good evidence is therefore partly a matter of its continued relevance to the situation as it stands now – and how it will stand in the ***future***, when (for example) responses are implemented. This all relates to the significance of evidence: what does the available evidence actually tell us?

So one key challenge is the generation of timely evidence and the renewal and testing of the evidence base against the evolving reality of a situation. Considering the other criteria for evidential quality listed in Section 2, ***accuracy*** is largely a function of methodological validity and sound implementation of surveys or other data-gathering methods (e.g. by trained enumerators using consistent methods etc.). This often poses substantial practical as well as theoretical challenges. The reliability of sources and observer bias as well as the ***representativeness*** of information are all recurrent problems that are best tackled either by statistical methods (such as sampling), or – in qualitative analysis – by triangulation of different information sources, cross-checking of data for consistency and by having multiple 'observers'. All of these again pose practical challenges in the time- and access-constrained context of an on-going crisis. Working with the 'best available evidence' of whatever quality in such contexts may be a legitimate, indeed necessary approach (Bradt 2009). But the question then arises: what have we done to generate the best available evidence?

From a policy-making perspective, the challenges to generating good evidence are somewhat different. Some feel that current evaluation practice generates little or no 'evidence' at all, in the sense of providing valid, evidence-based conclusions – for example, about the effectiveness of particular programme approaches – that can provide a sound basis for policy formulation. In the jargon, both the ***internal*** and ***external validity*** of the conclusions reached by evaluations are questioned. Other means of generating the kind of evidence that might provide a sound basis for policy-making – such as longitudinal analysis or controlled trials – are of debated application in the humanitarian sector. But generally speaking, there is a growing sense of need for more rigorous research, trials and systematic reviews in the sector.[10]

The generation of evidence comes at a cost in terms of money, time, opportunity and sometimes harm. As Bradt (2009) argues, 'Data-gathering and consequent humanitarian interventions are invasive

[9] Gerring 2012.

[10] See for example DfID 2012 strategy paper.

procedures with unintended consequences. Good intentions do not excuse bad outcomes.' Apart from opportunity costs and potential harm to affected communities, a particular problem in terms of cost-efficiency relates to the value of evidence beyond the agency that generates it. In many cases evidence is of little or no value to outsiders because it is not shared (Mills 2005), it is not shared in a timely manner (Darcy and Garfield 2011), it is not relevant to other stakeholders (Bradt 2009), or it is not in a format that is useful to other stakeholders (Poole and Primrose 2010). That said, there can be great value in independent surveys, which often serve as a wake-up call to the wider humanitarian community – and a prompt to action, including more systematic joint surveys.

In order to be usefully shared, evidence needs to be comprehensible to people other than those who generated it. At its most basic, this means that the language of documents needs to be accessible to outsiders. However, to make evidence more widely useful, there is generally felt to be an additional requirement for more standardised frameworks, indicators and methods of data collection. This is particularly significant in the conduct of needs assessment, in order for the level of needs to be compared within and across different contexts. Analysts have highlighted the need for a shared technical and conceptual language, and for common standards and methodologies for data collection (DFID 2012, p.31; Poole and Primrose 2010, p.1). However, the need for greater standardisation of methods and indicators should not detract from the importance of sharing information gathered by whatever means as the basis for joint discussion and coordinated action. In any case, mixed-method approaches, combining qualitative and quantitative data, depend heavily on interpretation. It is in the process of joint analysis and interpretation that much of the value of assessment evidence is found (Darcy and Hofmann 2003).

In the rest of this section we consider the ways in which evidence and knowledge are generated in practice in the humanitarian sector, and related questions about the quality and utility of available evidence. Our focus is on current practice in the sector across the formal diagnostic and learning processes, and particularly those which are standard elements of humanitarian programming:[11] evidence generated by early warning and surveillance systems, by needs assessment and monitoring

## The question of numbers

Some of the core quantities involved in 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 of which these numbers form part.

The lack of certainty about baseline population ('denominator') figures often remains unresolved throughout a crisis. But often the concern is with ***relative*** rather than absolute numbers, with (for example) percentages and rates that can be calculated in the absence of accurate overall population figures. Population sampling techniques enable representative data to be collected, demographic profiles to be established and the relative impact of disasters on different social groups to be assessed. Yet data disaggregated by age and sex remains the exception rather than the norm.

Refs: 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)

[11] Though our focus here is on evidence from formal processes, evidence gathered and knowledge acquired outside the formal processes may sometimes be of equal importance.

processes, and by project or programme evaluations. We consider the nature of evidence generated by each, particularly as judged against the 'strength' criteria set out in Section 2 above.

## 3.2 Evidence from early warning systems

**3.2.1** 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...'[12] 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 evidence is context- and situation-specific. It combines new data about a developing situation with historical knowledge to produce a predictive analysis of the likely outcome for a given area over a given timeframe. While the data concerned are context-specific, the analysis will draw both on 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, it depends for its force on being able to establish a convincing case for an emergent trend that left unchecked will lead to catastrophic outcomes (proposition type A above). Famine early warning is the best known example of this type. The case to be made is usually complex, combining data of different kinds, e.g. rainfall patterns, harvest yields, food prices, terms of trade, household income, malnutrition levels. Other cases may be simpler: early warning of an approaching cyclone, tracking its likely path; of impending flood; or even of an impending earthquake or volcanic eruption, though these remain much harder to predict.[13] Early warning potentially buys time to take preventive, preparatory or evasive action.

Effective early warning depends on the ability to project from an existing state of affairs to a likely future state (prognosis) based on causal analysis. It is thus closely related to ***risk analysis*** and ***scenario planning***. Here the availability and reliability of data may be at issue, and crucially so will the ***interpretation*** of available evidence (see further Section 4). DFID's 2012 strategy paper on innovation and evidence-based approaches asserts that decision-makers lack routine access to good information about risk. The approach proposed to address this involves risk modelling to inform resource allocation and programming, and standardised reporting of disaster losses (presumably to inform future risk models) (DFID 2012).

**3.2.2** Early warning is one of the areas of 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 made a major impact in reducing 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 been responsible for saving many lives.

**3.2.3** The evidence from most established early warning systems appears to score well for accuracy, when judged against the criteria for strength of evidence. But its probative value with respect to type A (diagnostic) propositions depends on the context. In rapid onset crises where (for example) cyclones and floods can be closely tracked, it is increasingly possible to say with a fairly high degree of certainty

[12] Source AlertNet: http://www.trust.org/alertnet/news/early-warning-of-disasters-facts-and-figures
[13] UNISDR 2006

whether the circumstances are likely to lead to catastrophic human impact and exactly which areas are likely to be affected. In cases of slow onset disasters like drought-related food security crises, it may be much more difficult. In these crises, multiple factors combine to determine the effect of such events on people and their livelihoods. The case to be made is complex, the evidence itself tends to be patchy and sometimes contradictory, and trends may not always be clear-cut. In such cases, it may be hard to show *ex ante* that a given set of early warning data demonstrates conclusively (or even with a high degree of certainty) that a crisis is imminent. Even where the evidence is strong, persuading institutions to take preventive action can be a major challenge. In Section 4 below, we consider the case of Somalia in 2011 as an example of where the lack of consensus over the probative value of available evidence led to serious delays in response to the emergent famine.

## 3.3 Evidence from needs assessment

**3.3.1** 'Needs assessment' describes a wide range of practices from informal observation and consultation with crisis-affected communities to formal, survey-based processes that may involve multiple parties and multiple sectors. The nature of the evidence that is generated varies widely depending on the methods adopted, which may in turn be dictated by factors like access and time constraints. Lack of access to affected areas at the onset of an emergency often makes it difficult to collect detailed information on needs. Furthermore, in insecure contexts, current data are often lacking and there are often additional political and resource constraints (Banatvala and Zwi 2000). In order to compensate for this, heavy reliance is often placed on existing knowledge of context. One evaluation concluded that 'analysis of humanitarian needs in Southern Sudan lacked in coverage and comparability of data during the war had been compensated for to a degree with nuanced and hard-won contextual analysis embedded within a cadre of committed and experienced humanitarian workers' (Poole and Primrose 2010, p.3). This suggests there may be some kind of substitutability between different kinds of evidence.

The current practice of needs assessment has been described as highly fragmented, patchy and flawed in a variety of ways.[14] Commentators have observed that assessments tend to be front-loaded (i.e. concentrated on the early first phase of a crisis), poorly documented, and potentially biased because they are conducted by agencies that are using them to bid for funding.[15] They tend not to be shared publicly, but rather to be used for internal programme design purposes and to substantiate funding requests. Where they are more public and inclusive, they are often either too slow to inform critical resource allocation decisions, or else they are too cumbersome, complex and compromised to provide clear or reliable evidence to guide action. There are exceptions to this of course, and considerable progress has been made in the development of more appropriate methods and tools of assessment.[16] Nevertheless, most individual agency practice on assessment has arguably not changed fundamentally in the past two decades. As a result, what should be one of the primary diagnostic tools of the sector is in practice not felt to be playing the role that it should.

One area where progress *has* been made is in the field of coordinated needs assessment – multi-agency and multi-sector. Needs assessment was recognised as one of the problems left unresolved by the UN's initial humanitarian reform agenda. The Inter-Agency Standing Committee (IASC) Sub-

[14] Bradt 2009

[15] Darcy and Hofmann (2003); de Ville de Goyet and Moriniere (2006); Darcy et al. (2012)

[16] To give one example, WFP now has a range of tools – from Comprehensive Food Supply and Vulnerability Analysis to Emergency Food Security Assessments and related market analysis tools – that can provide a considerably stronger evidential base for response to food security crises than was previously the case. The development of these tools followed a process of deliberation, testing and expert consultation over a number of years, funded by ECHO and other donors.

Working Group subsequently established a Needs Assessment Task Force headed by the Office for the Coordination of Humanitarian Affairs (OCHA), which has been active is developing common tools for the assessment of needs in rapid onset disasters. This aims to address one particular set of evidential problems relating to situational analysis: the need for rapid, accurate and multi-sectoral information to guide initial responses to rapid onset crises, and specifically to provide an evidence base for emergency Flash appeals. Another initiative – the Assessment Capacities Project (ACAPS)[17] – run by a consortium of INGOs, addresses a different set of needs. These concern the need for an independent 'read' on crisis situations and the supplementing of existing agency capacity with specialist assessment skills. ACAPS also produces 'disaster needs analyses' for particular crises based on secondary data analysis.

Garfield et al. (2011) analyse some of the perceived advantages and disadvantages to coordinated or 'common' needs assessments (CNAs) as currently practised. Among the potential advantages they list efficiency, coherence and coordinated planning and action. But they see potential disadvantages: slowness, expense, and the problems of reconciling different approaches to analysis. They also note a tendency to concentrate on more easily comparable quantitative, survey-based data at the expense of qualitative data. While the advantages are felt to outweigh the disadvantages, it is clear that joint needs assessment answers only some of the challenges to providing timely and reliable evidence for decision-makers.

It is important to mention here one other area of progress – or at least an area of new potential for evidence gathering. This lies in the use of new technologies and techniques of 'crowdsourcing' using social media as part of the assessment and monitoring of crisis situations and the evaluation of humanitarian responses. Two short examples must suffice here. One concerns the use of social and news media in tracking the course of disease outbreaks. In a review of the 2010 cholera outbreak in Haiti, reports on Twitter and news websites were found to correlate well with official government statistics – and were available up to two weeks earlier.[18] The second example concerns the recent use of mobile phones to consult affected communities in Haiti and Somalia, as part of both the needs assessment and evaluation processes.[19] Both examples illustrate the potential utility of such technologies in communicating with, and getting real-time feedback from, people in affected communities in a way that was previously difficult or impossible to achieve. This potential has yet to be fully exploited.

**3.3.2** The concept of 'demonstrating need' is not a straightforward one,[20] and what constitutes 'evidence of need' may depend on the observer. In particular, the external view may be at odds with that of those actually experiencing the crisis. The extent to which the views of crisis-affected people are used as evidence to inform response is variable. As DFID (2012, p.31) observes, it is 'important that we consider what kind of evidence counts. The experiences of disaster-affected communities

[17] An initiative of HelpAge International, Merlin and Norwegian Refugee Council. See http://www.acaps.org/
[18] Chunara, R. et al. (2012) 'Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak'. The authors conclude: 'During infectious disease outbreaks, data collected through ... official reporting structures may not be available for weeks, hindering early epidemiologic assessment. By contrast, data from informal media are typically available in near real-time and could provide earlier estimates of epidemic dynamics.'
[19] IASC RTEs for Haiti and Somalia; Chunara, R. et al. (2012) 'Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak'.
[20] As Darcy and Hofmann point out (2003), 'assessing needs' is an ambiguous concept. If 'need' is taken to mean a deficit or gap of some kind (as it tends to be), particularly a gap in goods or services, then responding to 'need' invites supply-driven responses aimed at filling that gap. This interpretation also depends on a logic that suggests that need does not arise until a catastrophic deficit occurs, which fails to explain the humanitarian case for preventive action. The authors propose an alternative view of needs assessment based on risk and outcome analysis: 'need' on this view is better understood as 'what needs to happen [by way of intervention] to avert catastrophic outcomes and promote preferable ones'.

are a rich source of evidence both of need, and the relative effectiveness of interventions across the humanitarian cycle. Experience in collecting this sort of evidence is increasing, but there is a strong need to systematically involve beneficiaries in the collection and use of data to inform decision making.' As noted above, new technologies are making feasible new methods of consultation with affected populations on an on-going rather than just a one-off basis.  Yet although practice has improved in this respect since the damning verdict of the Indian Ocean Tsunami evaluation,[21] the experience of the Haiti earthquake in particular suggests that there is far to go.[22]

Needs assessment is as much about understanding people's priorities, coping mechanisms and normal practices as it is about data collection and technical analysis. Ethnographic and anthropological methods of investigation and analysis may be required to understand the behavioural dimensions of a given situation. How the (generally qualitative) material that is generated by such methods is read alongside other data (particularly quantitative), depends on the skill and judgement of those collecting and interpreting it. Some of the more highly developed methodologies, such as the Household Economy Approach to measuring food and livelihood security,[23] have well-established methods for combining data of different kinds.

**3.3.3** Evidence from monitoring and surveillance systems is an essential adjunct to needs assessment, though many feel that it is under-resourced (Darcy 2009). Surveillance systems are most developed in health and nutrition, and tend to combine regular data collection with more in-depth surveys. So for example, Médecins Sans Frontières (MSF) Belgium runs a disease surveillance system in the Democratic Republic of Congo involving a network of 'antennae' that act as the first stage in a wider

## Biases, errors and the presentation of evidence

One of the biggest threats to the accuracy and credibility of evidence arises from ***bias***. Needs assessment, particularly as it is currently practised, is subject to a range of potential biases. The incentive of the assessing agency has already been noted as one potentially major source of bias, where that agency is using the assessment to support a request for funding of its own activities. But there are other less obvious forms of bias. Observer bias is a recognised hazard in any observational research. It arises when a researcher's own cognitive bias (preferences, assumptions, preconceptions, etc.) causes the researcher to influence the course of the trial or to interpret information arising from it in certain ways. This may be quite independent of any organisational incentives.

In statistical terms, bias is defined as any form of systematic (non-random) error – which means not a mistake but a ***deviation*** from the expected or 'true' result. Inclusion and exclusion errors involve the tendency to err either towards including or excluding people when identifying a target population. The more usual tendency in the humanitarian sphere is to err on the high side – risking inclusion rather than exclusion errors in the beneficiary list. The reasons for this in a given case may be quite justified, but it can lead to serious distortion and any such tendency should be acknowledged in the presentation of data. More generally, if the strength of evidence is to be properly assessed, it should be presented in ways that are as far as possible transparent about potential bias and error. So for example, statistical estimates should state confidence intervals as a

[21] Tsunami Evaluation Coalition (2006)
[22] Grunewald et al. (2010)
[23] See Seaman et al. (2000)

diagnostic system. Where a disease outbreak is reported, a 'ground-truthing' assessment (survey) is subsequently launched to confirm the truth of the report and obtain more detail. In Ethiopia, the nutrition surveillances system is a joint effort of government, international and national agencies, designed to spot malnutrition 'hotspots' which act as a guide to the targeting of nutritional support programmes.

'Monitoring' in this context more often refers to ***programme*** monitoring. For Oxfam, this is an 'on-going process carried out during programme implementation', which ideally includes a baseline study against which data on indicators of change collected on field visits can be measured, to be supplemented by mid-term and end-of-programme evaluations.

**3.3.4** Needs assessments often combine situational analysis with response analysis: in the terms of this study, they involve gathering evidence to formulate propositions first of type A, then of type B (and sometimes then of type C). So for example, a WASH (water, sanitation and hygiene) assessment may determine that 10,000 people lack access to clean water following a flood that has contaminated water sources. The assessment concludes that the consequence will be a public health disaster unless there is intervention. The same assessment may conclude that the temporary provision of trucked water will ensure access in the short term while water sources are being rehabilitated, thereby protecting public health. The relative feasibility, appropriateness and cost-effectiveness of this and alternative options may form part of the same assessment. The assessment is then (typically) used to substantiate a proposal for funded intervention – although the proposal and response plan are typically written by different people from the ones who conducted the assessment.

In terms of evidence strength, the results of needs assessments are highly variable. Those based on

### The practical limits of diagnostics

A doctor may make a differential diagnosis of disease that includes a range of possible causal explanations of the presenting symptoms, in the absence of conclusive evidence about cause. Given the uncertainty, she may begin by treating only for the most serious possible cause and outcome – i.e. in our terms, her Proposition B is based on a ***worst case*** or urgent Proposition A. If the patient subsequently recovers, the doctor may not know whether this was because the disease responded to the treatment (or what exactly the disease was) or because of some other self-correcting process. She may not think it worth investigating further: diagnostics cost time and money, and can sometimes cause harm. Since Proposition A no longer holds true (the patient has recovered), it may no longer be pressing to find out the nature of the disease and why the patient recovered.

Such uncertainty over causes and outcomes is common in the humanitarian context. If, for example, an outbreak of diarrhoeal disease in a refugee camp diminishes following a range of emergency public health measures, the assumption may be made that this was caused by the measures taken. But as Bradt (2009) points out, the assumption may not always be warranted. Commenting on similar claims made about the effectiveness of interventions in response to a hepatitis outbreak in Darfur in 2004, he notes that 'there are not enough data to demonstrate ***causation*** … Association is not causation and the agencies' response probably had little to do with [the decline in hepatitis cases]', which followed the expected epidemiological path. This is not to say that the measures taken were the wrong ones; simply that their efficacy in this case could not be demonstrated with the available data.

established methods of data collection and analysis – such as cluster sample surveys – may have high levels of reliability, other methods less so. Certainly their credibility may depend on the method of data acquisition. The extent to which assessment evidence is ***significant*** will depend in large part on extent to which inferences can safely be drawn from the indicators measured. For example, if 60 per cent of households surveyed said that adult family members had reduced the number of meals they consumed over the past month from three to two each day, what would that demonstrate? It is likely that only by combining a range of indicators can a solid proposition about food insecurity be formulated, raising important questions for the design of assessments.

## 3.4 Evidence from evaluations and controlled trials

**3.4.1** Evaluations of humanitarian action are faced with similar challenges to needs assessments and monitoring. They often take place in data-poor, politicised and complex environments, where physical access is limited, populations are mobile, and there are a variety of different actors all of whom wish to legitimate their view of what happened. In these contexts, evaluators face a variety of evidential tests as set out in Section 2.4.2 above. In particular, they need to build an evidence base which is accurate, representative of the experience of the affected population, and not biased by their own subjective interpretation. In this context, much of 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 utilize the advantages of both' (IFRC 2011, p.35). Similarly, MSF's evaluation unit suggests that 'usually a mix of qualitative and quantitative methods provides the best results' (MSF 2012, p.7), while WFP's guidelines suggest that 'as qualitative and quantitative data complement each other, both should be used' (WFP n.d p.23).

However, in practice, humanitarian evaluation 'uses mainly qualitative methods' (Buchanan-Smith and Cosgrave 2012). A review of evaluations in the ALNAP Evaluative Resource Database (ERD) suggests that use of mixed methods approaches are uncommon, and that the majority of evaluations  have tended to rely on 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, of 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.

It is also worth noting that, despite the heavy reliance on interviews as a source of evaluative data, many evaluations lack 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 only included 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.

**3.4.2** In addition to the other evidential tests set out in Section 2, evaluators need to address the challenge of causality and ***attribution***. It is not enough for an evaluation to accurately depict a situation; it needs also to show the relationship between a specific intervention, or series of interventions, and the situation described (i.e. proposition type B). As a result, evaluations need to be rigorous in their approach to attribution and, if the results are to be held to be valid beyond the specific context of the evaluation – that is, if they are to be used for the generation of policy – they also need to be externally valid (i.e. generalisable).

Many evaluations of humanitarian action address the 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 (decrease in water-borne diseases).[24] 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 between the deliverables and the outcomes that discounts alternative explanations, then the deliverable is generally held to have caused the outcome.

This approach, while arguably imperfect, has tended to dominate where 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[25] tends to be more complex and ambiguous. 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; and 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 these reasons militate against robust impact assessment (Proudlock and Ramalingam 2009). As a result, evaluations of impact are still fairly infrequent in the humanitarian sector (ibid.) and, where they do occur, the meaning of 'impact' often differs from one agency to another (ALNAP 2012; ACF, Mimeo). Notwithstanding this, there has recently been a significant increase in interest in humanitarian impact evaluation, driven at least partly by a desire to establish which approaches offer best value for money in a constrained financial environment.

### 3.4.3 Controlled trials and experimental approaches

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 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 US$ 200,000 to US$ 900,000 (World Bank n.d, quoted in Bradt 2009) – noting that cheaper, non-experimental approaches are regularly used in 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 n.d.). Critics also maintain that RCTs are of limited use in policy-making because they cannot be generalised to other settings (Schwandt 2009) and because they seldom explain how an intervention led to specific impacts (Piccioto). We can perhaps conclude that RCTs are certainly useful when answering the specific type of 'PICO' clinical question[26] for which they

[24] The need to test this logic against reality is apparent from the example given in s.4.2.3 below.

[25] 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).

are designed (Bradt 2009). 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' (Victoria et al. 2004, quoted in Dijkzeul et al.).

In the humanitarian arena, there have been some attempts to use RCTs to provide evidence of impact. Action Against Hunger (ACF) has conducted RCTs around nutrition programmes in Chad; DFID has funded an 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 the International Rescue Committee (IRC) has conducted an impact evaluation of a Community Driven Reconstruction programme in Liberia using an experimental design (Fearon et al. 2008). The IRC work is an interesting – and so far 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

### Research designs for investigating attribution

**Experimental design:**
In an experimental design, participants are randomly assigned, either to a group that receives programme services or to a control group that does not receive these services. The control group 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 randomly assigning participants to groups. Instead, they compare outcomes for groups who receive services and for similar groups who did not receive services (a 'natural experiment'); or for one group before and after an intervention.

**Theory-based approaches:**
These approaches do not attempt to demonstrate attribution by comparison of recipient group and counterfactual. 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'.

Refs: Definitions based on Stern et al. 2012; Morra Imas and Rist 2009; Leeuw 2012.

[26] 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 particularly relevant for issues such as determining which therapy will be most effective.

of evidence generation in the humanitarian sector.

At the same time, there will continue to be situations where experimental approaches are not possible, not desirable, or not feasible. Both Stern, in a recent paper for DFID (Stern 2012) and Rogers, in a paper for Interaction (Rogers 2012) suggest a variety of 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 the intervention. The Emergency Capacity Building project (ECB) has developed and tested a methodology that considers the contribution of humanitarian interventions to change, using descriptive statistics combined with interview data. The Feinstein Centre has produced guidance on participatory impact assessment which notes the practical and ethical difficulties of establishing control groups to test attribution and instead focuses on using participatory tools to assess the relative contribution of project and non-project factors to change (Catley et al. n.d).

# SECTION 4 – HUMANITARIAN DECISION-MAKING AND THE USE OF EVIDENCE

## 4.1 Background

**4.1.1** In this section we consider the ways in which evidence and knowledge are used by humanitarian decision-makers, the extent to which they are accessible to those who need it, and whether they play a central or peripheral role in the decision making process. We make comparisons with the use of evidence in other sectors, and consider this in the context of decision-making theory and organisational incentives.

**4.1.2** In Section 2 above, we suggested that there were three main types of humanitarian proposition requiring evidential support, concerning diagnosis, effective response and appropriate response. Evidence is important in defining, testing and refining these propositions, in convincing decision-makers and other stakeholders of their validity, and in the making of decisions. Agencies must convince donors that a crisis exists, that the proposed intervention will be effective in averting its worst aspects, that this intervention is the best option available for addressing the crisis, and that the agency concerned is capable of implementing that intervention to agreed standards. Agency staff have to convince their managers of the strength of particular evidence and the validity of particular or general propositions about a humanitarian response. In some cases, it may take little to convince the party concerned. In other cases, it may take much more: for example, where the crisis is not recognised or the context is not a strategic priority; where the proposed intervention is a novel or unproven one; where resources are scarce; or where the proposing agency is itself untested.

**4.1.3** Evidence is often used selectively, depending on the interests or priorities of the person or organisation in question. Indeed policy may be formulated despite the evidence of 'what works'. In some cases, evidence is made to fit the policy rather than vice versa, and in other cases evidence (e.g. from evaluation findings) is ignored or not shared because the findings are politically sensitive (Guenther et al. 2010).

**4.1.4** Individuals in different positions are often in possession of – and may favour – different types of evidence in relation to humanitarian propositions. As a result they may draw different conclusions as to the validity of a given proposition. For example, an agency manager in headquarters may prioritise evidence from global studies on the risk of malnutrition, while agency staff in the field may privilege evidence gained from their own interactions with the affected population. Neither type of evidence is 'wrong', nor is one necessarily stronger than the other in supporting (or challenging) a particular proposition. The point is that different actors need to be convinced of the strength of different kinds of evidence in support of different propositions, and decision-makers need to assimilate and use that evidence in making their decisions. This is difficult because no source of knowledge is infallible, all evidence must be interpreted, and the different sources of evidence on which our judgement lies are often not commensurable (Hammersley 2005). Thus 'knowledge from personal experience and from new research evidence must each be evaluated in its own terms, and then combined in some way that takes account of their distinctive characteristics as sources of knowledge' (ibid. p.88).

**4.1.5** The timing of evidence may be crucial to its uptake by decision-makers. Reviewing the uptake of a rapid initial assessment of needs by ACAPS following the Haiti earthquake, Darcy and Garfield (2010) 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.' More generally, we might conclude that evidence has to be communicated in a way that is timely if it is to be useful – and that having imperfect or raw data is preferable to having none at the point of decision-making.

### Typology of humanitarian decision-making

An ODI discussion paper on the use of information in crisis response decisions proposed a typology of four main decisions relating to crisis response in the humanitarian sector (presented here in slightly modified form):

- **Strategic** decisions about whether and how to respond, including macro resource allocations (approach/modality, level and channel of funding, etc.)
- **Programme design** decisions (including targeting)
- **Planning** and micro resource allocation decisions: what resources (money, people etc.) to allocate and how to allocate them (team composition, budgeting, etc.)
- **Operational** decisions concerning programme implementation and modification.

The study also distinguished levels at which decisions were made:

(i) Within organisations: HQ, regional, national, local/field levels
(ii) System-wide or inter-organisational.

Ref: ODI (2009) *Humanitarian Diagnostics: the use of information and analysis in crisis response decisions* (Discussion paper commissioned by FAO).

**4.1.6** The use of evidence may be expected to vary between different types and levels of decision-making. The Overseas Development Institute (ODI) paper cited in the box below suggests that the greatest concentration on evidence appears to be at the programme design stage. Strategic decisions are perhaps the most liable to be made with reference to 'external' factors like previously agreed strategic priorities. The extent to which operational decisions are informed by evidence depends in part on whether there are effective feedback loops from the programme implementation level to programme managers.

**4.1.7** Producing compelling evidence about a situation is only part of the battle. For example, it may be possible to show that the prevalence of global acute malnutrition in a given region has doubled from 5 to 10 per cent over a certain timeframe. But that does not in itself make a compelling case for action, or tell us what that action should be. Indeed, different people may draw different conclusions from the same evidence. Similarly, even the most compelling evidence about the impact of particular interventions may not provide a clear guide as to the appropriate response in a given case. [27]

## 4.2 Use of early warning evidence

**4.2.1** 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, such as to be able to formulate type A propositions with some confidence in many contexts. The use made of such information has been more problematic, particularly with regard to famine and food insecurity. The problem of disconnect between

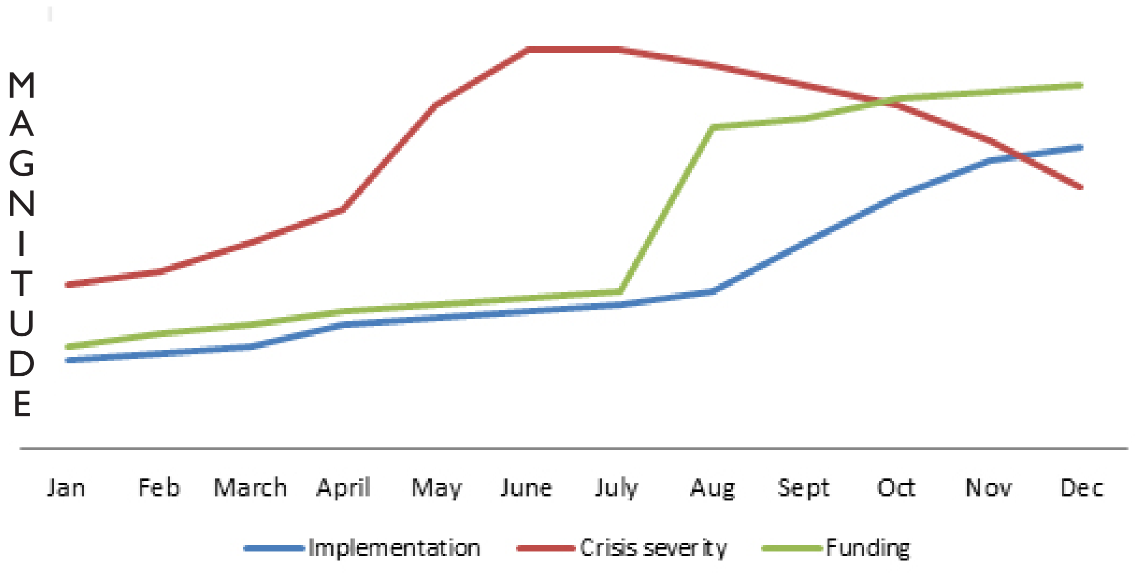


**Figure:** Schematic representation of response to the Somalia famine of 2011
[Source: IASC Real Time Evaluation of the international response to the Somalia Crisis 2011 - Valid International, 2012]

[27] By way of alternative example, an exhaustive 10-year randomised controlled trial in the UK into the effects of badger culling on the incidence of bovine tuberculosis concluded that the net effect of culling after 10 years was a 16 per cent reduction in the incidence of tuberculosis in cattle in the culling areas. Although the results were widely accepted as valid, the pro- and anti-culling lobbies continue to argue as to the significance of the results and its implications for policy. What it did do, however, was to narrow the terms of the debate from what had been essentially an ideological dispute about culling to an argument revolving mainly around cost-benefit. Report at http://archive.defra.gov.uk/.

early warning information and response decisions has long been recognised (see e.g. Buchanan-Smith and Davies 1995) yet the problem remains at the heart of policy and decision-making in slow onset food crises (Levine et al., 2011).

**4.2.2** We illustrate some of the issues here with reference to the 2011 famine in south central Somalia. The diagram below, taken from the IASC real-time evaluation of the Somalia crisis response, is a schematic representation of the relationship between crisis severity, response funding and response implementation.

The most striking feature of this picture is the gap between crisis severity[28] and funding availability, and then between funding and response implementation. 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 refer to as 'a systemic failure of early response' which had 'two aspects:

- A failure of prevention action, to tackle the proximate causes of vulnerability through urgent livelihoods intervention, so building short-term resilience and reducing the need for relief;
- A failure of scaled-up early relief, tackling the most acute symptoms of the crisis at the time when such assistance was most needed in early to mid-2011' (Valid International, 2012, p.37).

The absence of an overall contingency plan for major crisis response meant that planning for a massively scaled-up response only really took place in July 2011. The authors note: '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.'[29] By the time programme responses were being scaled up, the crisis was already well past its peak.

It is striking in this case that the 'evidence' that really galvanised the international response was not from early warning mechanisms but from a combination of media images of severely malnourished Somali refugees arriving at camps in Kenya and Ethiopia, and the declaration of famine in south central Somalia by the UN, based on a dramatic shift in nutritional indicators. Given the extreme political and security constraints on operating in these largely Al Shabaab-controlled areas, it is not entirely surprising that the evidence threshold was so high in this case. Nor is it surprising that it took visual evidence of acute food insecurity (indeed famine conditions) and use of the term 'famine' to cut through the political factors involved, particularly for donors. Yet this pattern was largely repeated in the wider Horn of Africa. The peculiar circumstances of Somalia, in other words, do not explain the phenomenon.

**4.2.3** It seems that outcome indicators – in this case, mainly indicators of acute malnutrition and excess mortality – are taken as far stronger evidence for the existence of a crisis than risk or 'leading' indicators of the kind used by early warning systems. This raises fundamental questions about how 'crisis' is understood, and the kinds of evidence that have to be presented to make the case for preventive intervention. As the diagram above illustrates, the overriding problem with reliance on outcome indicators as triggers in a slow onset crisis is that they do not provide a basis either for acute preventive action or for timely relief. This in turn raises the question as to how a compelling, evidence-based case

[28] '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 the diagram is indicative only.
[29] Valid International (2012). A detailed analysis of the issues involved can be found in Global Food Security Vol. 1(1) special issue on the 2011-12 famine in Somalia. See also the 2012 report by Save the Children and Oxfam 'A Dangerous Delay'.

can be made for intervention before the emergence of critical outcome indicators. As Levine (2011) and others have argued, the answer lies partly in recognition of recurrent seasonal patterns and the cumulative effect of shocks and stresses on fragile livelihoods. While it may be hard to say where the tipping point lies given the multiplicity of causal factors, some situations (like Somalia in 2011) are so extreme that the only safe conclusion to reach is that the combined result of rain failure, food price increases and other factors will be catastrophic. In other words, while the outcome may be uncertain, both the likelihood and potential impact (i.e. the risk) of the events in question are such as to demand pre-emptive action.

## 4.3 Use of needs assessment and monitoring evidence

**4.3.1** Commentators have noted major barriers to the uptake of certain evidence by decision-makers, possibly in relation to the weaknesses of the evidence produced by some current needs assessment practice. Darcy et al. (2012, p.31) note that there 'appears to be a high level of "path dependence" in most decision-making processes in the sector. In other words, the range of options is limited by previously decided strategic priorities, resource allocation, and other factors.' These parameters are sometimes set by host government authorities; in other cases they are set more by donors and by implementing agencies. 'This significantly limits the extent to which decisions are open to influence by evidence, particularly where organizational incentives to generate and respond to new evidence are limited.' (ibid. p.32)

Decision-makers may be highly selective in their uptake and interpretation of evidence (Darcy et al. 2012). Personal biases, rules of thumb and mental models – as well as a variety of (dis)incentives – may prevent individuals and organisations from responding to a situation in the way that evidence appears to demand. 'It is common for experienced staff to base decisions mainly on past experiences, instinct, and assumptions – even in the face of contradicting evidence. In institutional terms, this in turn leads to building agency capacity around established intervention types, which continue to be the "preferred response" with each new crisis, irrespective of available evidence.' (ibid. p.32.)

**4.3.2** Even where documented assessments exist, the link between assessment and decision-making appears weak. Moreover, assessments are still largely front-loaded and used to justify proposals or appeals (Bradt 2009). As Darcy et al. (2012, p.32-3) note, 'it remains the case that most assessments are conducted in order to substantiate a case made for funding by a particular agency to do a particular

### The psychology of decision-making

When faced with complex problems or incomplete information, rather than undertake taxing calculations, people tend to resort to simple educated guesses, 'rule-of-thumb' thinking or personal intuition. Psychologists refer to these as 'heuristics' (see for example, Gilovich, Griffen and Kahneman 2002) or 'biases'. As noted above, these tend to shape individual decision-making in significant ways. One of the main challenges to promoting evidence-based decision-making is to overcome inherent biases and habits of thought, and to allow evidence to challenge an individual's normal assumptions. This relates to the subject of incentives: an individual who is encouraged and rewarded for grounding decisions in evidence (or indeed penalised for not doing so) is more likely to challenge their own instinctive responses and to seek out relevant information.

Source: Adapted from Darcy et al. (2012)

thing. Inevitable biases result in a lack of credibility – both of the analysis and of proposed interventions based on that analysis. This appears to be a major distorting factor in the system. It creates a potential incentive to exaggerate the trigger event and its impact in order to secure as much funding as possible for the whole duration of the emergency during the critical "window" at the outset of the crisis'.

**4.3.3** Situational and outcome monitoring are essential complements to needs assessment, particularly given the potential distorting factors noted above. An example from the literature on project monitoring serves to illustrate the critical importance of testing project output-based 'theory' against the available evidence of outcomes. Diarrhoea is one of the five major causes of death in an emergency setting and one of the three main causes of death in children (Curtis and Cairncross 2003). An article published in *Disasters Journal* in 2005[30] concerns the response to an outbreak of diarrhoea in Abou Shouk camp for displaced people in Northern Darfur, Sudan. Although minimum standards were followed in the provision of water points and latrines, this was evidently not enough to prevent the spread of disease. The authors describe the knowledge base that informed the subsequent action:

> *It is now well recognised that the provision of clean drinking water at collection points is not enough to prevent water-borne diseases (Kaltenthaler and Drašar, 1996). Contamination often occurs while water is being collected, including from the handpump nozzles themselves (Clasen and Bastable, 2003), from the use of dirty containers, or during storage in the home (Mintz et al., 1995)… Although the washing of hands with soap is recommended as "the most effective measure to prevent transmission of Shigella" (WHO, 1995), as demonstrated by several field studies (Curtis and Cairncross, 2003), behavioural change strategies take time to implement… (p.214)*

While hand-washing was recognised as a vital component of hygiene promotion, 'a speedier intervention capable of generating instant results was deemed necessary for the Abou Shouk outbreak.' Testing of water sources ruled out the sources themselves as the origin of the infection. In the circumstances, it was decided to launch a campaign of mass disinfection of all water containers in order to break the contamination cycle. Diarrhoea figures from the clinics showed a fall in cases following the cleaning campaign, although the authors note that it is 'extremely difficult to obtain good and statistically rigorous data in an emergency setting'.

This case is interesting for a number of reasons. First, it shows that it is never sufficient to rely on the delivery of outputs according to best practice (water points, latrines, etc.) to deliver outcomes. Second, outcome indicators – in this case clinic data on the incidence of diarrhoea – have to be monitored if the success of the intervention is to be gauged. Third, it shows the essential value of an enquiring attitude to programme delivery: is this working, and if not, why not? In this case, the team responsible for delivery was faced with evidence that the measures put in place were not sufficient, and they launched an investigation into the reasons. This drew on prior knowledge (personal and literature-based) as well as direct observation. In the end, the team focused on the most likely proximate cause of infection – contamination of containers – and treated for that. This was and remained a hypothesis, though the subsequent fall in the incidence of diarrhoea strongly suggested that they had correctly identified and eliminated at least one main contributory cause.

The constraints of real-world factors on ideal practice are apparent from this example. The team had to find a way of quickly tackling the problem, and so focused their efforts on immediate rather than

[30] Walden et al. (2005) 'Container contamination as a possible source of a diarrhoea outbreak in Abou Shouk camp, Darfur province, Sudan.' *Disasters 29(3)*: 213–221.

longer-term solutions. In research terms, the example has limited value. As the authors note (p.215): 'The intervention presented here was not planned as a research study to measure efficacy; it was simply carried out to stop the diarrhoea outbreak.' It was only after the intervention was completed that the authors considered it interesting enough to be written up. Hence, there are obvious gaps in terms of both the data collected and information about the situation.

The authors of the above study raise the issue of the ethics of research in humanitarian contexts. In this case, 'a control group would have been preferable, but this raises ethical issues with respect to research of disease outbreaks in IDP camps.'. Although the ethics of research are not discussed here, it is of course a significant topic in considering how evidence is generated and used.

## 4.4 Use of evaluation evidence

**4.4.1** The number of humanitarian evaluations has grown significantly over the past decade. The ALNAP Evaluative Resource Database (ERD) – 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 work) 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, p.108).

Although the number of evaluations has been steadily growing over the last decade, 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 practice and 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);
(ii) Improving a programme (e.g. on-going learning and development);
(iii) Generating knowledge.

As Sandison notes, both uses (i) and (ii) are 'intended to lead to direct changes and decisions. This expectation of use is often referred to as "instrumental": an evaluation's findings and recommendations should lead to related actions such as tangible changes in policy, funding, systems or operational practice. Many – perhaps most – humanitarian evaluations fall into this category of instrumental use. Evaluations commissioned by donors at the end of a programme or partnership cycle, audits, mid-term reviews, real-time evaluations and so on may have different users and emphases but they share the same expectation of utilisation. They all assess merit, identify strengths and weaknesses and provide recommendations on what to do as a result.' (p.3)

In the terms of this paper, the first two kinds of use are likely to have a lower evidential threshold than the third. The evidence for instrumental purposes has to be persuasive at least, but since the related propositions make no claim to general (external) validity, it does not have to be conclusive. The third category is different, and the extent to which it is taken as contributing to knowledge beyond the institution concerned may depend on the rigour, independence and perceived overall validity of the evaluation process itself.

### 4.4.2 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, p.91). She adds, 'instrumental use is the least likely form of utilisation' (ibid., p.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., p.91).

In fact, while it is not hard to find examples of evaluation recommendations that have been ignored, and while '[i]n general the literature describes an inconsistent and, in some cases, a dismal record of evaluation use' (ibid.), the picture is by no means wholly negative. Professional evaluators contacted during the preparation of this paper consistently pointed to recommendations that had been implemented. For example, the report of the TEC 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.

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 per cent of recommendations were implemented (OCHA 2007). When WFP studied the degree to which evaluation recommendations had been taken up, they found that 54 per cent had been implemented and 65 per cent had been included in successor documents (WFP 2005). WFP found that recommendations were more likely to be implemented where they were operational, rather than strategic, and where their implementation only required action from a limited number of people. Broader recommendations, or those which required coordination with partners or headquarters units, were less likely to be implemented, as were recommendations with intangible benefits or those which implied criticism of WFP staff (ibid.).

All of which suggests that the evidence provided by humanitarian evaluations is frequently 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.

**4.4.3** Over the past decade, many humanitarian organisations – including DFID, SIDA, UNICEF and WFP – have attempted to identify the ways in which they can improve the uptake and use of evaluative evidence. ALNAP has also published three papers on the topic (Hallam 2011; Sandison 2005; van de Putte 2001) based on these experiences and those of other Network members. Among other recommendations, 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 for 'follow-up' exist.

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 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).

**4.4.4** 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 which 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 explicitly tied 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 the Humanitarian Country Team in the first instance, a move which may be intended to increase country ownership of the results.

In Haiti, Groupe URD have implemented what Grunewald calls 'Real Time Evaluation plus'. Here, a team conduct a series of evaluations of the same project over a period of two years, and in the process, work closely with the project team. The later evaluations concentrate largely on identifying progress made with the recommendations of the previous missions and identification of new challenges. As Grunewald 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' (Grunewald 2012, para. 3).

ACF is also encouraging dialogue between evaluators and field staff, in an attempt to increase the utilisation 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 earning is significantly strengthened (Guerrero 2012; Allen 2012).

### 4.4.5 Evidence and policy – the use of evaluations and research

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, has led to broad acceptance of cash programming, and the agency recently lifted the 100,000-euro 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, p.24). USAID has recently changed its Food For Peace Title II policy to explicitly include cash transfer programming. 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 US$ 74.9 million in 2006 to US$ 188.2 million in 2010 (Global Humanitarian Assistance 2012). Although funding subsequently fell in 2011, the general trend would appear to be for an increase in the use of cash. This is not least because large agencies are planning to significantly increase their activities in this area.[31] As a programming approach, 'cash-based work in humanitarian relief has shifted … from radical and risky to … mainstream' (Ramalingam, Scriven and Foley 2009, p.43).

To what degree has evidence played a role in this policy shift? A previous ALNAP study documented the evolution of the use of cash in humanitarian programming (ibid.), and the key points are worth repeating here. 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 methodically reviewed 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' (ibid., p.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 (ibid., p.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.[32] 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).

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, p.17), and made a significant contribution in focusing 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., p.22). Moreover, while some of the JEEAR research was influential, many of the most important conclusions 'were ducked and have been consistently evaded' (ibid., p.24). We consider one of the conclusions that did not lead to change below.

**4.4.6** 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

[31] Personal communication, Haley Bowcock, Cash Learning Partnership (CaLP) secretariat
[32] See http://www.cashlearning.org/what-we-do/research-focus

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, p.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, p.3). Policy development is not exclusively evidence-based, and evidence is not always used to develop policy. Two examples can perhaps illustrate the failure of the humanitarian system to make strategic or policy changes on the basis of evidence.

The first example is one of the other conclusions that came out of the JEEAR – one that appears to have been 'consistently evaded' (Buchanan-Smith 2005, p.25). 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' (JEEAR: Study 3 1996, p.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 which 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 who 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 organizations [and]; include at least some staff or advisors with considerable experience in the country' (JEEAR: Synthesis 1996, p.61).

**4.4.7** Over the next decade, these findings were echoed in an appreciable amount of academic research which 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 regularly returned to this theme (see for example Ali et al. 2005, Oxfam 2004, World Vision 2011, Nicholson and Desta 2010, Boku 2010). However, a decade later 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, p.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 (SOHS 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 (SOHS 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).

**4.4.8** This is not the only failure to create robust policy responses to situations where evidence suggests that change is required. A second 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, generally triggered by unacceptably high levels of malnutrition. However, over the last two decades, there have been a number of calls to move to an alternative 'early response' model, in which agencies respond to early warning of drought by 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' (p.33). Four years later, Jaspars, reviewing the literature and conducting case studies, came to similar conclusions (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 for example 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, did lead 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. p.6). Over the last decade, livelihoods interventions were 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 per cent of total funding for drought relief in 2011 (Catley 2012). In 2012, eleven 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 US$ 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, p.13).

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 (JEEAR; Redmond 2010). In the case of early interventions, Levine, Crosskey and Abdinoor have noted that the number of programmes, and therefore the evidence-base, remain limited (Levine, Crosskey and Abdinoor 2011).[33] 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 various factors militate against putting this knowledge to use. These factors include: 'the inflexibility and supply-driven nature of the international relief system' (JEEAR: study 3 1996, p.177); 'donor stipulated restrictions on how [agencies] use funds' (TGLLP Steering Cttee, p.11); and 'the urgency to spend money visibly' (Telford and Cosgrave 2006, p.93). There is probably more than a grain of truth in this. But again, it is only part of the story. Policy changes aimed at building long-term partnerships with local civil society actors, ensuring more locally recruited staff are in decision-

[33] 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 formally test the assumption – which raises the interesting question of the role of logic models in providing evidence.

making positions, or reducing turnover in emergencies, seem eminently possible and would go a long way to ensuring programmes were based on stronger contextual knowledge. Donors could release more funds earlier, and agencies 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, Scriven and Foley 2009, p.44)). So why, so often, does it appear 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 action on evaluations which 'challeng[e] strongly held beliefs and behaviour embedded in the organisation's culture' (Sandison 2006, p.111), and concludes that 'using evaluation is as much a people issue as a technical one' (ibid., p.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, p.45). Effective change – including the development and introduction of new policy – requires a process which addresses the rational, political and emotional needs of stakeholders in the organisation.

**4.4.9** In recognition of 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 seem to also be important in 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, Scriven and Foley 2009, p.3). Although cash programmes had been practised for some time, and in a variety of contexts, the 2000 IFRC report 'Buying power' was the first time that the results of these programmes became readily accessible to policy-makers. HPG, and latterly CaLP, have subsequently 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, p.22) and benefitted from a funded follow-up process which 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 partly result 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, p.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).

**4.4.10** It would appear that access to information is not, however, sufficient. 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 field level learning; Sandison 2007; Darcy 2009; Maxwell mimeo). 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' (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). This tendency (which has worrying implications for the ability of 'local' knowledge to influence the direction of humanitarian action) can be seen to have been influential in the acceptance of cash as a programming tool. The case study suggests 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' (Ramalingam, Scriven and Foley 2006, p.55), 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. 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, Crosskey and Abdinoor 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 (2011).

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., p.98). In the aftermath of the Rwanda response, for example, with agencies engaged in internal debates about how to improve and donors demanding action, these push factors were particularly strong, and this undoubtedly served to ensure 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 factors 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 militate 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. Meanwhile, donors 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, Crosskey and Abdinoor 2011). Donors, too, are often incentivised to spend larger sums of money than the NGOs' request for livelihoods work; Levine, Crosskey and Abdinoor 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, p.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).

Given these constraints, it will be interesting to see whether the experience of the 2011 famine in south central Somalia, and the increasing popularity of livelihoods approaches, lead to any significant policy changes. It is interesting, too, to consider the external conditions which might lead to increased consideration of operational contexts in humanitarian programming – and whether a combination of increased assertiveness from affected states and the ability of disaster-affected people to use new media to broadcast their experiences of aid might lead to a 'context revolution'.

### Use of evidence in the nursing sector

Perhaps more than any other, the nursing sector has engaged in critical debate as to the value and role of evidence (as understood in the scientific tradition) in decision-making. This is particularly interesting because nursing shares some characteristics with humanitarian action, and these characteristics have been salient in shaping the debate. Specifically, nurses and humanitarians share a duty or ethic of care. Further, modern nursing is patient-centred, emphasises the importance of patient dignity, and seeks to enable and empower patients to participate in decisions concerning their own care and treatment. To varying degrees, ideas of beneficiary dignity, empowerment and participation are also expected to inform humanitarian response.

The move to evidence-based nursing has been met with strong criticism from elements within the nursing centre. Two strands of critique may be particularly pertinent to thinking about the use of evidence in the humanitarian sector. The first critique relates to the need to combine professional experience and judgement with evidence of the effectiveness of particular treatments. Thus a broad definition of evidence is advocated, and should include 'expert knowledge, clinical experience, patient perspectives, stakeholder consultation, evaluation of previous policies, non-experimental research and other secondary sources' (Kitson 2002, p.180).

The second retains a narrower definition of evidence, but argues that 'evidence-based practice obstructs nursing process, human care, and professional accountability' According to this perspective, the nurse-person is absolutely central to nursing practice, and human relationships are not best directed by the results of experimental research. Indeed such an approach is deemed to be 'inconsistent with professional ethical codes, with current philosophical thought, and with what people say they want from nurses... the nurse-person process is not data based – it is human based and must be guided by values and theoretical principles' (Mitchell 1999, p.32).

# SECTION 5 – SUMMARY OF CONCLUSIONS

## 5.1 General conclusions

**5.1.1** The analysis presented in this paper suggests that the humanitarian sector has made some progress in grounding its practices and policies in evidence, but that this progress is at best inconsistent and is in many respects weak. This raises the question of whether and how current practice might be strengthened. We have reviewed some of the lessons that might be applied more widely from inside and outside the sector, and the factors that might limit their application. We conclude here by summarising the main points arising from the discussion above and by setting out some of the key questions arising for the sector.

**5.1.2** The demand for more evidence-based practice and policy in the humanitarian sector has grown in recent years, partly as a result of donor pressure. This demand concentrates particularly on two areas. The first concerns the extent to which analysis of needs and proposed responses in any particular crisis is grounded in evidence (from needs assessment, established best practice, etc.). The second is more generic and concerns the building of the knowledge base and the accumulation of evidence to inform the development of policy. These two strands are linked but distinct.

**5.1.3** Any discussion of evidence has to be clear on the question: evidence for what? We suggest that such a discussion in the humanitarian sector needs to distinguish at least three different types of proposition to which evidence is applied: those concerning the problem statement or 'diagnosis' of crisis situations (A); those concerning the effectiveness of a given response option (B); and those concerning the choice in practice between alternative responses based on appropriateness, feasibility, value for money and other criteria (C).

**5.1.4** Each of these proposition types requires different kinds of evidence to substantiate them. But a common feature of all three is that the preferences and attitudes of crisis-affected people must be factored into analysis; and mixed methods of enquiry (qualitative and quantitative) will almost always be required to gain a true picture of what is happening. Neither a purely 'scientific' focus on what can be measured and quantified, nor exclusive reliance on perceptions and subjective feedback, would ideally be used to inform decision-making. Moreover, we should not focus on establishing evidence around any one particular area of specialist concern – e.g. food security, health or nutrition – in isolation from another.

In short, the evidence base for humanitarian action is necessarily diverse in nature. This raises some difficult methodological problems about how to combine different types of data, particularly across the quantitative-qualitative divide. It puts an onus on good methodological design, but also on the skills and judgement of those interpreting the evidence.

**5.1.5** We tentatively propose five criteria for testing the quality or strength of evidence: truth (or accuracy); representativeness; significance; generalisability; and (the validity of) causal attribution. But we recognise that there are other ways of judging evidential strength, not least according to the source of the evidence and the way it was generated, e.g. from controlled trials or observational studies, cluster sample surveys, 'expert' or local opinion, and a range of other more or less formal ways of assessing needs and response options.

**5.1.6** Turning to current practice, the quality of evidence from assessments and evaluations is highly variable. Agency biases cast doubt both on the accuracy of their situational assessments and on the appropriateness of their proposed responses, as well as on the generalisability of findings. Other biases may be more subtle and individualised. The accuracy and relevance of assessments is also potentially compromised by the fact that they tend to be conducted at the onset of a crisis, and are not consistently followed up – so that it is unclear whether their findings hold true as the crisis evolves. Monitoring and surveillance systems are often weak or non-existent. More generally, the technical quality of assessments is variable and the results are often not shared with other agencies. The use of new communication technologies and social media hold the prospect of being able to communicate with and get feedback from affected communities throughout the course of a crisis, but this has yet to be exploited to anything like its full potential.

**5.1.7** Evaluations of humanitarian responses tend to rely almost exclusively on qualitative methods and on purposive sampling to identify interviewees, while often failing to include beneficiary groups among those consulted. They rely heavily on logical frameworks and related inferences in assessing causal links between interventions and outcomes – and tend to have little to say about impact. The use of more experimental approaches to gathering evidence – and particularly the use of randomised controlled trials – is growing. However, experience from other sectors, and particularly the development sector, suggests that the applicability and utility of RCTs will be debated. We should expect their use to grow in those contexts where it is both ethical and feasible to conduct them, but RCTs seem likely to remain the exception rather than the norm in the sector as a whole. Elsewhere there would appear to be plenty of scope for more rigorous and systematic use of quasi-experimental approaches, case-based approaches, observational studies and meta-evaluations. However, as these alternative approaches to evaluation have not been used to any great extent in the humanitarian sector, their applicability remains to be tested.

**5.1.8** Generating good quality evidence is not enough. Evidence has to be used. A variety of factors influence the degree to which evidence is used in making decisions. The ***timeliness*** and ***relevance*** of evidence to particular kinds of decision can have a major bearing on the degree to which the evidence is considered by decision-makers. How that evidence is ***presented*** or ***communicated*** is also likely to have a significant effect. But it seems that multiple factors may limit the uptake and use of such evidence. Some relate to the biases and assumptions of individuals; others to more organisational factors, including established ways of working and previously determined priorities – themselves sometimes shaped by political factors.

**5.1.9** There are some specific usage challenges relating to particular kinds of evidence. The use of information from early warning systems relating to slow-onset disasters is fundamentally hampered by a lack of consensus about what constitutes a humanitarian crisis that would demand urgent action. The humanitarian system is far more geared towards responding to ***outcome*** indicators than to risk indicators, with the result that it is almost inevitably 'behind the curve' in responding to emergent crises. The use of needs assessment and monitoring data is held back partly by issues of quality and availability of data, partly by the high degree of path dependence built into decisions about response.

The use of evaluation evidence is dominated by expectations of ***instrumental*** use: i.e. the belief that evaluation conclusions should directly inform decisions about a particular organisation's policies, systems and practices. In practice, instrumental use does not always occur, although evaluations are more likely to be acted upon where the conclusions are more specific and operational. The practices of Real Time Evaluations (now required by the IASC in all level 3 emergencies) and iterative evaluation involves more substantial consultation with programme staff than traditional 'independent' evaluations, and are closely tied into decision-making processes.

The use of evaluations as a means of generating more generic evidence and knowledge is much less emphasised. Many are simply not designed to yield such generalisable evidence. However, some evaluations do aim to produce generalisable evidence that can be used to inform policy, and in doing so, add to the growing body of more formal research into humanitarian issues. Where such evidence is generated there remains a series of political and institutional obstacles to strategic change, especially where there is no great external pressure for change.

**5.1.10** To what extent is the job of interpretation of evidence the domain of 'experts', and to what extent is it the domain of programme decision-makers and policy-makers themselves? Put another way, what can the generalists reasonably expect from the specialists? One thing they might reasonably expect is that their specialist colleagues talk to each other, share evidence, and make recommendations about response (types B and C) that are properly informed by an understanding of other sectors and of key cross-cutting factors, i.e. social (including gender and age), economic, political, security, etc. Ultimately, however, it is the job of decision-makers on the 'implementation' side to ensure they are properly informed, to assimilate the available evidence – and to ask questions where evidence is lacking.

## 5.2 Questions for discussion

This paper has attempted to explain why humanitarians are increasingly interested in the topic of evidence and knowledge; to provide some definitions of knowledge, evidence, and evidence 'quality'; and to investigate current practices in the generation and use of evidence in the humanitarian sector. In doing so, it provides a background for discussions at the 28th ALNAP Annual Meeting. Our review of current practice in the generation and use of evidence in humanitarian action raises some important questions. These include:

### General

- If the humanitarian sector is not sufficiently evidence-based in its practice, to what extent is the problem one of lack of ***availability*** of (good) evidence, and to what extent is it lack of proper use of available evidence? What are the main challenges under each of these headings?

### Generation of evidence

- How 'fit for purpose' is the evidence currently generated from formal diagnostic and evaluative systems, i.e. baseline analysis, early warning, surveillance, needs assessment, situational and programme monitoring, as well as various forms of evaluation?
- Do our assumptions about evidence affect the degree to which affected people can influence humanitarian operations?
- How does the evidence currently produced score when judged against criteria of truth (accuracy); representativeness; significance; generalisability; and (validity of) causal attribution?
- Is the humanitarian sector over-reliant on either qualitative or quantitative evidence at different stages of the project cycle?
- Do humanitarians have the necessary skills to generate high quality evidence?
- Is the generation of data – and the process of evaluation – adequately built into programme design from the outset? What should be the relationship between programme implementation and impact monitoring?
- What are the relative roles of programme functions (such as needs assessment and evaluation) and research in generating evidence for policy. How can both be enhanced?
- Should there be more use of experimental research methods in the humanitarian sector? If so, what are the priorities for such research?
- At what point does 'diagnostic' or 'learning' investigation cease to be cost- or time-effective?

## Use of evidence

- Is it possible to agree on a common performance criterion related to the use of evidence? E.g. *Was the best available evidence used to inform the response?*
- What is the proper role of evidence in decision-making? How, for example, does evidence relate to individual judgement and to political imperatives?
- How can risk (leading) indicators – as distinguished from outcome (historic) indicators – form a more convincing basis for early action in slow-onset or protracted crises?
- How can we ensure that evidence is made available for decision-makers?
- How can decision-makers best balance different types of evidence?
- What are the implications of complexity and non-linear causality in the humanitarian context for the use of evidence and the kinds of evidence required?
- Does the 'instrumental' use of evidence for programme adjustment compromise the search for more robust evidence to support more general propositions and help build the evidence base for the sector as a whole? Should the emphasis be shifted?

# BIBLIOGRAPHY

**Abebe, D., Cullis, A., Catley, A., Aklilu, Y., Mekonnen, G. and Ghebrechirstos, Y. (2008)** 'Livelihoods impact and benefit-cost estimation of a commercial destocking relief intervention in Moyale district, southern Ethiopia,' in *Disasters 32(2)*: 167–86.

**ACF (2011)** *Learning review.* Available at: http://reliefweb.int/sites/reliefweb.int/files/resources/Situation_Report_71.pdf.

**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*, OUA IBAR. Feinstein International Famine Centre, School of Nutrition Science and Policy, Tufts University.

**Ali, D. et al. (2005)** 'Cash Relief in a Contested Area: Lessons from Somalia'. *HPN Network paper no. 50.* London: Overseas Development Institute, Humanitarian Practice Network, March 2005

**Ali, D. (2012)** 'A deadly delay: risk aversion and cash in the 2011 Somalia famine,' in *Humanitarian Exchange, issue 54*, HPG/ODI, London.

**Allen, B. (2012, September 6)** 'Collecting and reporting best practices from the field' [Msg 1, thread 2 under Capacity Area 2: Evaluation policy and purpose]. Message posted in ALNAP Evaluation capacities - Community of Practice (private forum).

**Ashdown, P. (2011)** *Humanitarian Emergency Response Review.* Available at: http://www.dfid.gov.uk/what-we-do/key-issues/humanitarian-disasters-and-emergencies/how-we-respond/humanitarian-emergency-response-review/

**Banatvala, N. and Zwi, A. B. (2000)** 'Public health and humanitarian interventions: developing the evidence base' in *BMJ 321*: 101-5.

**Bekele, G. and Abera, T. (2008)** *Livelihoods-based Drought Response in Ethiopia: Impact Assessment of Livestock Feed Supplementation.* Pastoralist Livelihoods Initiative.

**Bekele, G. (2010)** *Review of Save the Children US's Livestock Marketing Initiative in Southern Ethiopia.* Save the Children USA.

**Beynon, P. et al. (2012)** *What difference does a policy brief make?* Institute of Development Studies and the International Initiative for Impact Evaluation (3ie).

**Bhattacharjee, A. and Lossio, R. (2011)** *Evaluation of OCHA Response to the Haiti Earthquake.* ALNAP ERD. New York.

**Bickman,L. and Reich, S. (2009)** 'Randomised Controlled trials: a gold standard with feet of clay?' in Donaldson, S. Christie, C. and Mark, M. (2009) *What counts as credible evidence in applied research and evaluation practice.* Thousand Oaks, CA: Sage.

**Boku (2010)** *Participatory Impacts Assessment of Drought Reserve Areas in Guji and Borana Zones, Oromia Region.* Report prepared for Save the Children USA.

**Bradt, D. A. (2009)** 'Evidence-based decision-making in humanitarian assistance'. *HPN Network Paper 67.* London: Overseas Development Institute, Humanitarian Practice Network.

**Broadbent, E. (2012)** 'Politics of research-based policy in African policy debates'. *Synthesis of case study findings.* Evidence-based Policy in Development Network.

**Buchanan-Smith, M. and S. Davies (1995)** *Famine Early Warning Systems and Response: the Missing Link.* London: IT Publications.

**Buchanan-Smith, M. (2000)** *Role of Early Warning Systems in Decision-Making Processes.* London: Overseas Development Institute, Humanitarian Policy Group.

**Buchanan-Smith, M. and Cosgrave, J. (2013)** *ALNAP Guide: Evaluation of humanitarian action.* Unpublished manuscript.

**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 3)*. London. Available at: www.alnap.org/

publications/7RHA/Ch3.pdf

**Buchanan-Smith, M. and Scriven, K. (2011)** *Leading Effectively in Humanitarian Operations.* London: ALNAP. Available at: http://www.alnap.org/resource/6118.aspx.

**Buchanan-Smith, Margie (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, J., Hovland, I. and Young , J. (eds) *Bridging Research and Policy in Development. Evidence and the Change Process.* London: ODI. Available at: http://www.odi.org.uk/rapid/Publications/BRP_ITDG.html.

**Burns, John C, et al. (2008)** *Impact Assessment of the Pastoralist Survival and Recovery Project, Dakoro, Niger.* Lutheran World Federation and Feinstein International Center.

**Catley, A., Burns, J., Abebe, G. and Suji, O. (n.d.)** *Participatory impact assessment: a tool for practitioners.* Boston MA: Feinstein International Centre, Tufts University.

**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.* Available at: http://sites.tufts.edu/jha/archives/1548

**Clarke, P. and Ramalingam, B. (2008)** 'Organisational change in the humanitarian sector' in *ALNAP 7th Review of Humanitarian Action.* London: Overseas Development Institute.

**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.

**Currion, P. (2010)** *Coordination at the Crossroads: NGO coordination in Southern Sudan 2007-2011.* Available at: http://www.alnap.org/pool/files/co-ordination-at-the-crossroads-ngo-coordination-in-southern-sudan.pdf

**Darcy, J. and Garfield, R. (2011)** *Review of the Rapid Initial Needs Assessment for Haiti*, paper prepared for ACAPS. Geneva: Assessment Capacities Project.

**Darcy, J. and Hofmann, C-A. (2003)** *According to need? Needs assessment and decision making in the humanitarian sector.* London: Overseas Development Institute.

**Darcy, J. et al (2012)** *The Use of Evidence in Humanitarian Decision Making.* ACAPS Operational Learning Paper. Boston: Tufts University.

**Development Initiatives (2012)** *Global Humanitarian Assistance: Tracking spending on cash transfer programming in a humanitarian context.* Wells: Development Initiatives.

**de Ville de Goyet, C. and Moriniere, L. C. (2006)** *The Role of Needs of Assessment in the Tsunami Response.* London: Tsunami Evaluation Coalition.

**DFID (2012)** *Promoting innovation and evidence-based approaches to building resilience and responding to humanitarian crises: a DFID strategy paper.* London: DFID.

**DG ECHO (2009)** *The Use of Cash and Vouchers in Humanitarian Crises: DG ECHO funding guidelines.* Available at: http://ec.europa.eu/echo/files/policies/sectoral/ECHO_Cash_Vouchers_Guidelines.pdf

**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. doi:10.1111/j.1467-7717.2010.01187.x

**Dijkzeul, D., Hilhorst, D. and Walker, P. (2012)** 'Evidence based action in humanitarian crises'. Manuscript submitted for publication in *Disasters*.

**Epstein, R. M. et al. (2004)** 'Communicating evidence for participatory decision making' in *JAMA: The Journal of the American Medical Association, 291(19)*: 2359-66.

**Evans, I. et al. (2011)** *Testing treatments: better research for better health care.* London: Pinter and Martin.

**Fearon, J. Humphreys, M. and Weinstein, J. (2008)** *Community Driven Reconstruction in Lofa County: Impact assessment.* Available at: http://www.columbia.edu/~mh2245/FHW/FHW_final.pdf 10 January 2013.

**Featherstone, A. (2011)** *Strength in numbers: a global mapping review of NGO engagement in coordinated assessments.* Emergency Capacity Building Project.

**Feinstein International Center (2007)** 'Impact Assessments of Livelihoods-based Drought Interventions' in Moyale and Dire Woredas, *Ethiopia: A Pastoralist Livelihoods Initiative report.* Produced in Partnership with CARE, Save the Children US and USAID Ethiopia.

**Foresti, M. (2007)** 'A Comparative Study of Evaluation Policies and Practices in Development Agencies'. Report for AFD Evaluation Department.

**Goldacre, B. (2012)** *Bad pharma: how drug companies mislead doctors and harm patients.* London: Fourth Estate.

**Gottwald, M. (2010)** *Competing in the humanitarian marketplace: UNHCR's organizational culture and decision-making process.* Geneva: UNHCR Policy Development & Evaluation Service.

**Grünewald, F. (2012, November 5)** '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 (private forum).

**Grünewald, F. et al. (2010)** *Inter-agency real time evaluation in Haiti: 3 months after the earthquake.* Available at: http://ochanet.unocha.org/p/Documents/Haiti_IA_RTE_1_final_report_en.pdf

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

**Guerrero, S. (2012, May 18)** 'Our vision of linking evaluations and learning: ACF Learning Review 2011' [ALNAP blog post]. Blog posted at http://www.alnap.org/blog/72.aspx

**Hallam, A. (2011)** *Harnessing the power of evaluation in humanitarian action.* ALNAP, ODI. Available at: http://www.alnap.org/pool/files/evaluation-alnap-working-paper.pdf

**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: A Journal of Research, Debate and Practice 1(1)*: 85-100.

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

**Henry, G. T. and Mark, M. M. (2003)** 'Beyond use: understanding evaluation's influence on attitudes and actions' in *American Journal of Evaluation 24(3)*: 293-314.

**Hilhorst, T. (2003)** 'Responding to disasters: diversity of bureaucrats, technocrats and local people' in *International Journal of Mass Emergencies and Disasters, 21(3)*: 37-55.

**House, S. (2012)** 'Evaluation of Pakistan flood response 2011/12: using Oxfam GB's global humanitarian indicator tool'.

**HPG. (2006)** 'Saving lives through livelihoods: critical gaps in the response to the drought in the Greater Horn of Africa'. *HPG Briefing Paper.* Available at: http://www.odi.org.uk/sites/odi.org.uk/files/odi-assets/publications-opinion-files/2041.pdf

**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.

**IASC (2011)** 'Humanitarian System-Wide Emergency Activation: definition and procedures', PR/1204/4078/7, Geneva: IASC

**IFRC (2011)** 'Project/programme monitoring and evaluation (M&E) guide'. Available at: http://www.ifrc.org/Global/Publications/monitoring/IFRC-ME-Guide-8-2011.pdf

**Jaspars, S. (2006)** 'From Food Crisis to Fair Trade: Livelihoods analysis, protection and support in Emergencies', *ENN* special supplement series, no 3.

**JEEAR: Synthesis (1996)** *The International Response to Conflict and Genocide: Lessons from the Rwanda Experience: Synthesis Report.* Available at: http://www1.oecd.org/derec/sweden/50189495.pdf

**JEEAR: Study 3 (1996)** *The International Response to Conflict and Genocide: Lessons from the Rwanda Experience Study 3: Humanitarian Aid and Effects.* Available at: http://www.oecd.org/derec/50189439.pdf

**Johnson, K. 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.

**Jones, H. (2012)** 'Promoting evidence-based decision-making in development agencies'. London, Overseas Development Institute. Background Note.

**Jones, H. and Mendizabal, E. (2010)** 'Strengthening learning from research and evaluation: going with the grain', report to IACDI. London: Overseas Development Institute. Available at: http://www.odi.org.uk/publications/5154-

learning-research-evaluation-dfid?id=5154&title=learning%3B-research%3B-evaluation%3B-dfid.

**Kitson, A. (2002)** 'Recognising relationships: reflections on evidence-based practice' in *Nursing Inquiry 9(3)*: 179-86.

**Leeuw Frans L. (2012)** *Theory Based Evaluation.* Available at: http://ec.europa.eu/regional_policy/information/evaluations/pdf/impact/theory_impact_guidance.pdf (accessed 28 January 2013);

**Levine, S., Crossley, A. et al. (2011)** *System Failure: Revisiting the problems of timely response to crises in the Horn of Africa.* London: Overseas Development Institute.

**Maxwell, D. and B. Watkins (2003)** 'Humanitarian Information Systems and Emergencies in the Greater Horn of Africa: Logical Components and Logical Linkages' in *Disasters 27(1)*: 72-90.

**Maxwell, D., J. Parker, and H. Stobaugh. (2012)** What Drives Program Choice in Food Security Crises? Examining the "Response Analysis" Question. *World Development*, Special Edition on Impacts of Innovative Food Assistance Instruments, forthcoming.

**Mays, N. 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.

**Mazurana, D. et al. (2011)** *Sex and age matter: improving humanitarian response in emergencies.* Medford, MA: Feinstein International Center, Tufts University.

**Mills, E. J. (2005)** 'Sharing evidence on humanitarian relief' in *BMJ 331*: 1485-86.

**Minear, L. (2002)** *The Humanitarian Enterprise: Dilemmas and Discoveries.* Bloomfield: Kumarian Press.

**Mitchell, G. J. (1999)** 'Evidence-based practice: critique and alternative view' in *Nursing Science Quarterly 12(1)*: 30-35.

**Morra Imas, L. and Rist, R. (2009)** *The Road to Results: designing and conducting effective development evaluations.* The International Bank for Reconstruction and Development, World Bank: Washington.

**MSF Vienna Evaluation Unit (2012)** *Evaluation Manual: A handbook for Initiating, Managing and Conducting Evaluations in MSF.* Available at: http://evaluation.msf.at/fileadmin/evaluation/files/documents/resources_MSF/msf_evaluation_manual_2012_final.pdf

**Nicholson, N. and Desta, S. (2010)** 'Evaluation of the Enhanced Livelihoods in Mandera Triangle and Southern Ethiopia 2007–2009', ELMT/ELSE, funded by USAID.

**OCHA (2007)** *Management response matrix to the Intermediate Review of the Central Emergency Response Fund.* Available at: https://docs.unocha.org/sites/dms/CERF/Response_Matrix.pdf (accessed 10 January 2013).

**ODI (2009)** *Humanitarian diagnostics: the use of information and analysis in crisis response decisions.* London: Overseas Development Institute, Humanitarian Policy Group.

**OECD-DAC (2002)** *Glossary of key terms in evaluation and results based management.* Available at: http://www.oecd.org/findDocument/0,2350,en_2649_34435_1_119678_1_1_1,00.html (accessed 10 January 2013).

**Oxfam (n.d.-a)** *How are effectiveness reviews carried out?*

**Oxfam. (n.d.-b)** *Monitoring, evaluation, accountability and learning* [online]. Available at: http://policy-practice.oxfam.org.uk/our-work/methods-approaches/monitoring-evaluation (accessed 24 November 2012).

**Oxfam (n.d.-c)** *Oxfam GB evaluation guidelines.*

**Oxfam (n.d.-d)** *Rough guide to monitoring and evaluation in Oxfam GB.*

**Oxfam (2004)** *Evaluation of Oxfam GB's Food Aid and Food Security Emergency Intervention in Mauritania.* Acacia Consultants Ltd

**Pantuliano, S. and Wekesa, M. (2008)** 'Improving Drought Response in Pastoral Areas of Ethiopia, Somali and Afar Regions and Borena Zone of Oromiya Region', for the CORE group (REGLAP/ELSE/ELMT), HPG/ODI, January

**Peppiatt, D., Mitchell, J. and Allen, P. (2000)** *Buying Power: The use of cash transfers in emergencies.* London: The British Red Cross.

**Piccioto, R. (2012)** 'Probing the paradox of the RCT craze in international development', guest blog. Available at: http://ngoperformance.org/2012/05/24/guest-blog-robert-picciotto-on-randomised-control-trials/ (accessed 10 January 2013).

**Poole, L. and Primrose, J. (2010)** *Southern Sudan: Funding according to need.* Wells: Development Initiatives.

**Proudlock, K., Ramalingam, B., and Sandison, P. (2009)** 'Improving humanitarian impact assessment: Bridging theory and practice' in *ALNAP Review of Humanitarian Action (Chapter 2)*. London: ALNAP.

**Ramalingam, B., Scriven, K., and Foley, C. (2009)** 'Innovations in international humanitarian action' in *ALNAP Review of Humanitarian Action (Chapter 3)*. London: ALNAP.

**Ravallion, M. (2009)** 'Evaluation in the practice of development' in *The World Bank Research Observer 24(1)*: 29-53.

**Ravallion, M. (2011)** 'Knowledgeable bankers? The demand for research in World Bank operations'. *World Bank Research Working Paper No.5892*. Washington, DC: World Bank.

**Redmond, Anthony D., 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, December 2011.

**Robson, L. S. et al. (2001)** *Guide to evaluating the effectiveness of strategies for preventing work injuries: how to show whether a safety intervention really works.* Cincinnati, OH: Centers for Disease Control and Prevention; National Institute for Occupational Safety and Health.

**Rogers, P. J. (2009)** 'Learning from the evidence about evidence-based policy. Strengthening evidence-based policy in the Australian Federation' in *Roundtable Proceedings, Vol.: 1.* Canberra: Australian Government Productivity Commission: 195-213.

**Sackett, D., W. Rosenberg, et al. (1996)** 'Evidence based medicine: what it is and what it isn't' in *BMJ 312(5)*: 71–2.

**Sadler, K., Kerven, C., Calo M. et al. (2009)** *Milk Matters: A literature review of pastoralist nutrition and programming responses.* Feinstein International Center, Tufts University and Save the Children.

**Sandison, P. (2006)** 'The utilisation of evaluations' in *ALNAP Review of Humanitarian Action 2005*. London: Overseas Development Institute, pp. 89-144.

**Save the Children and Oxfam (2012)** 'A Dangerous delay: The cost of late response to early warnings in the 2011 drought in the Horn of Africa'. *Joint Agency Briefing Paper.* Oxford, UK: Oxfam International and Save the Children.

**Schwandt, T. (2009)** 'Towards a practical theory of evidence for evaluation', in Donaldson, S. Christie, C. and Mark, M. (2009) *What counts as credible evidence in applied research and evaluation practice.* Thousand Oaks, CA: Sage.

**Scriven, M. (2009)** 'Demythologising Causation and Evidence', in Donaldson, S. Christie, C. and Mark, M. (2009) *What counts as credible evidence in applied research and evaluation practice.* Thousand Oaks, CA: Sage.

**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: 17-31.

**Shaxson, L. (2005)** 'Is your evidence robust enough? Questions for policy makers and practitioners' in *Evidence & Policy 1(1)*: 101-12.

**Shaxson, L. (2012)** 'Expanding our understanding of K*(KT, KE, KTT, KMb, KB, KM, etc.)'. A concept paper emerging from the K* conference held in Hamilton, Ontario, Canada, April 2012 (Draft for comment). In: Bielak, A. et al. (eds.) Hamilton, ON: UNU-INWEH.

**Silverman, W.A. and Sackett, D. L. (1999)** *Where's the evidence? Debates in modern science.* Oxford: Oxford University Press

**Spencer, L. et al. (2003)** *Quality in qualitative evaluation: a framework for assessing research evidence.* London: Government Chief Social Researcher's Office.

**Stern, E. et al. (2012)** 'Broadening the range of designs and methods for evaluations: report of a study commissioned by the Department for International Development'. *DFID Working Paper 38.* London: DFID. Available at: http://www.dfid.gov.uk/r4d/pdf/outputs/misc_infocomm/DFIDWorkingPaper38.pdf

**Telford, J. and Cosgrave, J. (2006)** *Joint evaluation of the international response to the Indian Ocean tsunami: Synthesis Report.* Available at: http://reliefweb.int/sites/reliefweb.int/files/resources/F48164952D0AE1F0492571B700230406-tec-tsunami-14jul.pdf

**Tsunami Evaluation Coalition (2006)** *Joint evaluation of the international response to the Indian Ocean tsunami: Synthesis report.* London: ALNAP.

**UNHCR (2010)** *UNHCR's evaluation policy.* Geneva: UNHCR Policy Development and Evaluation Service.

**UNICEF (2004)** *UNICEF evaluation report standards.* New York: UNICEF.

**UNICEF (2007)** *Programme policy and procedure manual: programme operations* (revised February 2007). New York: UNICEF.

**United Nations Economic and Social Council (2007)** *UNICEF evaluation policy*, E/ICEF/2008/4.

**Valid International (2012)** *IASC Real Time Evaluation of the response to the Horn of Africa Drought: Somalia.* Oxford: Valid International.

**van de Putte, B. (2000)** *The Utilisation of Evaluation Recommendations in Medecins Sans Frontieres – Holland: A Study of 10 Evaluation Reports (1997– 1999).* Amsterdam: MSFH, May.

**van de Putte, B. (2001)** *Follow-up to Evaluations of Humanitarian Programmes.* London: ALNAP, April.

**Venton, C., Fitzgibbon, C., Shitarek, T., Coulter, L., Dooley, O. (2012)** 'The Economics of Early Response and Disaster Resilience: Lessons from Kenya and Ethiopia.' Economics of Resilience Final Report, funded by UKaid from the Department for International Development.

**VSF (2009)** 'Meat and Milk Voucher project (IMPACT I and II)', Clarke and Fison, VSF, Bahr al Gazal, South Sudan

**WFP (undated)** Monitoring and Evaluation Guidelines. Module: How to design a Results-Oriented M&E Strategy for EMOPs and PRROs. United Nations World Food Programme Office of Evaluation. Available at: http://documents.wfp.org/stellent/groups/public/documents/ko/mekb_module_8.pdf.

**WFP (2005)** *Summary report on WFP follow-up to recommendations.* Available at: http://documents.wfp.org/stellent/groups/public/documents/eb/wfp050928.pdf (accessed 10 January 2013).

**World Vision Pakistan (2011)** *End of Programme Evaluation Report for DEC-funded WV Relief Program in Sindh (July 2011- Phase 1).* Sustainable Solutions Ltd. Available at: http://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: Overseas Development Institute. Available at: http://www.odi.org.uk/publications/159-bridging-research-policy-international-development-analytical-practical-framework.

**Zhang, D. 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.