Chapter 3

The utilisation of evaluations

Peta Sandison

3.1 Introduction

Although evaluations are generally successful in identifying lessons and building institutional memory, only a minority of evaluations are effective at introducing evident changes or improvements in performance. If this continues, there is a danger that continued poor utilisation will undermine the credibility of evaluation as a tool for accountability and learning in the humanitarian sector.

We do not know even how many evaluations are conducted, let alone how many are used. The source of concern regarding non-use in the sector is mostly anecdotal – from working observations by evaluation managers and users – and proxy, from the apparent lack of impact of evaluation on the sector's performance. After all, if evaluations are intended to improve performance, then 'the recurrence of many of the problems seen in Rwanda and other emergency responses' observed in the synthesis report of the Tsunami Evaluation Coalition (Telford and Cosgrave, 2006, p23) implies that evaluations are not doing their job.

Evaluations of humanitarian action were rare in the 1980s. Twenty years on, the number has soared and evaluation has become one of the most visible features of the learning and accountability agenda. The growth continues, complemented by innovations such as joint and real-time evaluation, participatory evaluation and peer review. We have become increasingly interested in evaluating our performance, but are we using what we find?

The utilisation of research and evaluation has been a topic of lively debate in the development and public sector since the 1970s. ALNAP's humanitarian membership has long been concerned about utilisation, first commissioning a study on the follow-up to evaluations five years ago (van de Putte, 2001). The concern persists.

The results of studies by humanitarian agencies and donors on the use of their evaluations are mixed. Descriptions of limited or absent use are more easily found than examples of good practice. In general the literature describes an inconsistent and, in some cases, a dismal record of evaluation use. A Sida study concludes that 'evaluations are useful to a very limited group of stakeholders. For a majority of stakeholders the evaluation process could just as well have been left undone' (Carlsson et al, 1999, p51).

But others are more upbeat. WFP's study on its follow-up to evaluations carried out over a two-year period concludes that 88 per cent of the recommendations had been implemented or were in the process of implementation. Some two thirds of the recommendations had led to improved performance, although the reviewer recommends caution, noting that 'management units were almost always assessing performance intuitively' (WFP, 2005, p7). A 2002 study on the use of evaluations in the European Commission found that, while the degree of use varied, 'in no case were evaluations considered to be not at all useful' (Williams et al, 2002, p12).

The picture of utilisation that emerges from this and other studies is complex and often unexpected. One of the few certainties is that how and why an evaluation is carried out significantly affects the likelihood of it being used. The studies that constituted the background reading for this chapter provide a significant amount of information about factors that promote the utilisation of evaluations. Much of it resonates with van de Putte's 2001 study, and key references such as Michael Patton's *Utilization-Focused Evaluation*, published in 1997.

Information on use-promoting approaches is clearly available, but is the humanitarian sector using it? And if not, why not? Does the main issue continue to be the quality of evaluations (content and process), suggesting that the evaluation community itself is not learning? Or are there other issues that undermine even the best evaluations? Given the position of evaluation as a primary tool for accountability and learning, is it the right tool for the job? And what kind of job do we expect it to do?

3.1.1 Structure of this chapter

This chapter has five main sections. Following this introduction (Section 3.1), Sections 3.2 and 3.3 draw upon existing studies of utilisation to describe the different types of use made of evaluation, and examine a range of factors found to promote use. Section 3.4 examines four case studies and the findings of a series of interviews and a questionnaire survey. These findings are used to expand on the evidence of the preceding studies, exploring the extent to which factors affecting use identified in the literature are borne out in practice. Section 3.5 considers the implications for the future of learning and accountability mechanisms in the sector.

Box 3.1 Sources of information for this chapter

This chapter draws upon a literature review, four case studies of evaluation utilisation volunteered by CAFOD, MSF(H), OCHA and USAID, semi-structured interviews with 45 evaluators, evaluation managers and evaluation 'users', a review of 30 sets of terms of reference, and an electronic survey sent to ALNAP Observer and Full Members (19 evaluation managers and 27 evaluators responded to the survey).

The researcher was supported by an advisory group composed of representatives from the British Red Cross Society, CARE International, MSF(H), the Netherlands Ministry of Foreign Affairs (Policy and Operations Evaluation Department), OCHA and ODI, plus an independent evaluation consultant and the ALNAP Secretariat.

3.2 What is utilisation?

Utilisation in practice is complex and by no means static. It can mean different things to different users and is strongly related to context. In many cases, utilisation (or 'use' – the two terms are used interchangeably here) is impossible to measure and difficult to attribute to an evaluation.

3.2.1 Types of utilisation

Examples cited in the literature demonstrate a range of expectations of evaluation use. Patton (1997, p 76) describes three primary uses of evaluation findings:

1. judging the merit or worth of a programme (eg accountability to stakeholders; to inform funding decisions)
2. improving a programme (eg ongoing learning and development)
3. generating knowledge.

The first and second uses above are based on a cause-and-effect model. They are both 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. Utilising the evaluation means taking its advice.

The expectation of instrumental use is an understandably common one. Instrumental use makes good sense; a set of recommendations provides tangible targets that can, in theory, be measured and monitored. Instrumental use is coherent with the humanitarian sector's current emphasis on results-based management. It is, after all, reasonable to anticipate that the utilisation of an evaluation should be easily recognised and clearly linked to the findings. However, most studies of utilisation conclude that instrumental use is not the norm.

Since the 1970s, the neat, linear connection between evaluation (or research) findings and policy development or programme improvement has been increasingly challenged. 'Evidence suggests that government officials use research less to arrive at solutions than to orient themselves to problems' (Weiss, 1977, p534). Evaluations can influence the user's thinking and enhance understanding, contributing to the development of knowledge (Patton's third category of use, which is also often referred to as 'conceptual' use). Even if an evaluation's recommendations are ultimately rejected, the dialogue, dispute and debate over an evaluation's findings can generate increased clarity and new, more objectively grounded solutions. Such knowledge may or may not lead to action. Any action that does result is not direct and neither does it occur in the short-term. This type of knowledge-generating use is incremental and cumulative. It is typically associated with organisations and policy making, rather than with individuals and specific projects.

Evaluation syntheses might lend themselves best to conceptual use. Over time, the lessons learned and the repetition of key findings drawn from successive evaluations generate new information and subsequent dialogue about patterns of performance at an organisation- or sector-wide level. The utilisation of a synthesis

study is often a result of recognising recurring and therefore potentially systemic issues. The cumulative effect increases the status of the findings as a credible source of organisational knowledge, particularly for senior managers and policy-makers. In instrumental terms, the contributing evaluations may not have been used at all.

Unpredicted and unexpected uses, while not a separate analytical category of use, are certainly worth noting. Several examples were given of evaluations thought long forgotten, later being resurrected or actively used, without the knowledge of the authors or evaluation managers. This reinforces the general finding that the types of use that actually occur are not necessarily those that were planned, nor do they manifest in a predictable form.

Studies of use have also found that individuals and organisations can learn through the process of an evaluation, irrespective of the findings. An extensive review commissioned by the American Evaluation Association (Shulha et al, 1997) examined 74 studies of evaluation use. The review found an increasing prevalence and recognition of such 'process' use, such as participants' increased ownership of evaluation findings, greater confidence in the evaluation process and in applying the results and evidence of personal learning, all of which combine to produce a net subsequent effect on programme practice.

Patton (1997) further notes that users' participation can enhance an individual's commitment to learning and help build a stronger organisational learning culture, contributing to an organisational development process. According to Patton, a user's engagement in an evaluation can also lead to enhancement of shared understanding within teams and the development of monitoring and evaluation as an integral component of programme delivery. An example of this can be seen in a recent joint NGO evaluation in Niger when the evaluation process 'set a precedent for greater collaboration, sharing of resources and information and learning among agencies, and provides them with the opportunity to develop trust, and to regard one another not as competitors but as partners' (Wright and Wilson, 2006, p33).

Another type of use results when an evaluation confirms, rather than informs or creates, a position or knowledge that one or more key stakeholders already hold. This differs from misuse (which is covered below) in that the findings have not been manipulated or distorted to serve the user's purpose but have arrived independently at the same conclusion. A criticism of evaluations is that they often fail to discover anything new. This is particularly pertinent to field workers, frustrated at pages of

familiar information. For stakeholders seeking independent confirmation and the consolidation of information, this is not necessarily a problem. The evaluation is used as an objective source of evidence (to support, for example, the completion and closure of a programme), or as a means of communicating a stakeholder's position. The findings are already known to at least some of the key stakeholders.

A study of the utilisation of 10 MSF(H) evaluations (van de Putte, 2000) notes that the agency had often used evaluations to close an internal debate. Evaluations 'forced a decision'. In all cases, the findings corroborated the existing position of some stakeholders and, inevitably, lost the argument for others. The Sida study (Carlsson et al, 1999) observes that one of the evaluations was clearly intended to communicate the donor's concerns about gender and participation to its partner. The evaluation's subsequent identification of weaknesses in these areas was therefore nothing new to Sida, but it 'legitimised' its position and provided an independent reference for future discussions about programme approach, albeit in an expensive fashion.

Box 3.2 summarises these four principal types of use drawn from the literature, in particular from Carlsson et al (1999), Williams et al (2002) and Patton (1997).

Box 3.2 Types of use

Instrumental use Direct implementation of the findings and recommendations by decision-makers, leading to related decisions such as future funding, improvements to a programme or changes in policy and procedure. Evaluations that anticipate instrumental use include ex-post, accountability-judgement evaluations or audits and learning-improvement evaluations such as real-time or mid-term evaluations.

Conceptual use Evaluation results and conclusions trickle down into the organisation in the form of new ideas and concepts debated and developed over time. This type of knowledge-building use is sometimes referred to as 'enlightenment' use (Weiss, 1977). The effect is incremental: single evaluations rarely lead to direct changes in policy and practice but add to, clarify and develop knowledge. Conceptual use is also cumulative – findings may act as a reminder of

CONTINUED

Box 3.2 Types of use *continued*

knowledge. Conceptual use is also cumulative – findings may act as a reminder of what was known but previously put aside. Ex-post and evaluation syntheses naturally incline towards conceptual use. However, any type of evaluation or review can contribute relevant findings that generate cumulative knowledge.

Process (learning) use Participation in the evaluation itself can lead to individual learning and changes in behaviour, such as improved communication within teams and between partners, enhanced understanding and application of M&E in programming. Engagement in the process can also increase users' ownership of evaluation findings and the confidence to use them.

Legitimising use The evaluation legitimises – confirms, substantiates, corroborates – a decision or understanding that the organisation or individual already holds, providing an independent and objective reference that may be used to communicate or justify subsequent actions.

3.2.2 Misuse of evaluation

Non-use can occur for rational, unintended or practical reasons such as shelving a poor-quality evaluation, bad timing or unexpected events. Misuse, by contrast, is intentional and unrelated to the quality of the evaluation itself. Misuse occurs if an agency commissions an evaluation with no intention of acting upon it, or if an evaluation is carried out as a type of ritual, administrative routine or public-relations exercise.

Accountability is not satisfied simply by transparently identifying performance issues. It also involves a commitment to respond to the findings of an evaluation. Ritual evaluation means that 'an account' has been made (transparency), but no action follows. The number of such evaluations is unknown; if they are common then 'doing' evaluations has become more ingrained, while accountability has not. Such symbolic or ritual use of evaluations can be a cynical manipulation of an agency's public image. It may also be more banal; staff might simply be responding to their institution's administrative directives. The evaluation box is ticked. Either

way, money has been wasted and opportunity costs incurred through the diversion of staff time and energy.

Misuse also occurs when stakeholders deliberately attempt to subvert the independence and integrity of the evaluators and their findings. The Sida study (Carlsson et al, 1999) describes the 'hints' and subtle understanding that can be communicated to evaluators to serve a hidden agenda. While something of an exception in its refreshing honesty, Sida's experience of misuse is unlikely to be exceptional.

Box 3.3 Types of non-use or misuse

Ritual use Evaluations serve a purely symbolic purpose, representing a desirable organisational quality such as accountability. Evaluations are a formality and 'use' equates with the fulfilment of legal or institutional obligations, rather than of the evaluation findings.

Mis-use Suppressing, subverting, misrepresenting or distorting findings; co-opting evaluators to serve a biased agenda for political reasons or personal advantage. Rejecting findings because they do not correspond with the beliefs of key stakeholders or with decisions already taken (such as to cease funding).

Non-use The evaluation is ignored because users find little or no value in the findings (a rational response to a lack of quality or relevance), are not aware of the results (dissemination problem), or the context has changed dramatically (eg evacuation; unexpected closure of a programme).

These different uses and misuses are not mutually exclusive. Components of the same evaluation can lead to a mixture of uses at different times, partly related to the nature of the findings and partly to the users. Different users will select how, and if, they use the findings according to their position, power and interests.

3.2.3 Types of user?

An evaluation useful to headquarters staff in quest of organisational knowledge may not be valued by field staff seeking direct changes in their programme. Individuals within a team or department may also assess the utility of the same evaluation differently. A small number of chief executives were interviewed for this study to gain insight into how such key, high-profile stakeholders use evaluations. While only indicative at best, their descriptions are consistent with the types of use described above.

Box 3.4 Types of use by senior executives

- to support what you want or need to do: independent corroboration or a tool to influence (ie, the findings confirm the prevailing policy direction or individual agendas)
- an independent reference used to resist a directive from above (eg trustees, the board)
- as an outward demonstration of accountability
- to provide guidance when you do not know how to solve a problem
- as an advocacy tool
- as institutional memory.

The evaluation uses cited by the senior executives reflect the needs, interests and imperatives of leadership positions. The literature examined has not explored different types of use by a range of different users, although some logical assumptions could be drawn. For example, ALNAP's research on field-level learning (ALNAP, 2004) noted that field workers are more inclined towards learning from peers, rather than from knowledge contained in reports, guidelines and evaluations. The findings also indicated that field workers prefer learning by doing. If field

workers use evaluations at all, it could be expected that instrumental and process use would dominate, leading to immediate programme changes and learning through participation in the evaluation.

3.2.4 Defining and measuring utilisation

How an evaluation is used appears to defy reliable prediction and seems subject to many factors beyond the control of commissioning agencies or evaluators. The uses are also rich and more diverse than may be expected. If it is common that 'only direct instrumental use of findings and recommendations are regarded as "proper"; use' (Williams et al, 2002, p56), we are failing to recognise the many dimensions of utilisation and therefore doing evaluation a disservice. The picture would look a lot brighter if we accept the recommendation of the European Commission study that '[t]he indirect use of evaluations – including process use, indirect use and cumulative use – should be valued more explicitly' (Williams et al, 2002, p56).

However, such recognition of indirect evaluation use is a mixed blessing. How then do we define utilisation overall and, more importantly, how do we measure it? After all, a poor evaluation could escape censure on the grounds that its utilisation was non-instrumental – diffuse, indirect and therefore hard to discern. If we assess the performance of evaluation in terms of utilisation (as well as its conformity to good evaluation practice), how do we take account of genuine indirect use and learn how to serve it better?

If, for example, use equates with learning, then evaluation should be measuring its performance with respect to learning objectives. But measuring learning is a significant challenge. It is further complicated in evaluation because there is no baseline to compare with, the participants have not necessarily made any formal or informal contract with the evaluators to learn (such as making time for it), and detailed objectives of learning are rarely defined in the terms of reference (ToR) or in any explicit way during the evaluation. The evaluators themselves are often on a steep learning curve, absorbing vast quantities of information about the organisation, programmes, country and individuals and hence are sometimes poorly placed to judge how much of the information generated by an evaluation is new to the users or just new to them.

The difficulties are compounded by variable recall on the part of potential users. Weiss comments that 'the policy maker himself is often unaware of the sources of his ideas. Bits of information seep into his mind, uncatalogued, without citation' (p534). This is not to say that learning is not happening, or that 'learners' wilfully conceal their sources (although professional rivalry and defensiveness may sometimes account for this). Rather, the process of storing and recalling knowledge is complex, and its translation to action is a highly individual and personal process during which individuals transform knowledge and make it their own.

Furthermore, even measuring direct utilisation is problematic. It presupposes that the recommendations are the best or only solution and therefore should have been implemented. Humanitarian and development evaluations share the complexity of the work they are evaluating. Despite the increasing acceptance of sector-wide standards and protocols, evaluation, like programming, is at best a set of informed judgements made in contexts different from those of the events and activities being assessed. There is no gold standard. A dearth of impact data further leaves the door open to different opinions regarding the most effective approach, even in technical domains bound by more widely accepted indicators of good practice. As a result, the recommendations of even the 'best' evaluation can be disputed or rejected on perfectly rational grounds. This further confounds our ability to employ utilisation as a quantitative performance indicator for evaluation.

In summary, an evaluation leads to different types of use depending on the nature of the findings, the interests and needs of the users and the context. Use is neither wholly predictable nor objective. As we shall see in the following sections, there are many identifiable factors that can positively and negatively affect utilisation. Some, though by no means all, of these factors can be controlled or influenced by how an evaluation is designed and implemented. The presence or absence of many of these factors may provide a proxy method of assessing the performance of an evaluation in terms of its potential for utilisation. To do so requires a different definition of utilisation.

It is possible to define utilisation in qualitative terms related to an evaluation's *potential* for use, rather than struggling with definitions that exclude the richness of indirect use and ignore the complex mix of art and science that characterises evaluation use. Utilisation has occurred if the evaluation has led to relevant objective or subjective change as perceived by the users. Drawing on definitions found in

Patton (1997) and Carlsson et al (1999): an evaluation has been utilised if users with the intention and potential to act have given serious, active consideration to its findings, identifying meaningful uses according to their own interests and needs.

The next section explores the factors identified in the literature that promote, enhance and inhibit the use of evaluations.

3.3 Factors affecting utilisation

Why are some evaluations used while others are not? Studies on use have identified a number of influential factors. They can usefully be organised into distinct, though inter-related, groups to provide a framework through which we can analyse and perhaps predict the potential of an evaluation to be used.

The RAPID (Research and Policy in Development) Framework developed by ODI examines the importance of four dimensions that influence the impact of research on policy: the quality of the evaluation information, the political context, the links between the evaluators, policy-makers and other networks and, finally, the influence of the external environment. The RAPID Framework was applied to examine the influence of one of the most famously utilised humanitarian evaluations, the Joint Evaluation of Emergency Assistance to Rwanda (the JEEAR), on the development of the Sphere Project (Box 3.5).

The RAPID Framework provides a useful starting point for grouping the findings of the various studies on evaluation utilisation. For the purpose of this study, these have been grouped into four areas of factors affecting utilisation.

1 **Quality** of the evaluation process and product: the evaluation design, planning, approach, timing, dissemination and the quality and credibility of the evidence.

2 **Organisational** culture and structure: learning and knowledge-management systems, structural proximity of evaluation units to decision-makers, political structures and institutional pressures

3 **Relational** factors: relationships and links between the evaluators and users and their links to influential stakeholders, including shared backgrounds and common ground; relationships between key stakeholders; networks and links between the evaluation community or team and policy-makers

4 **External** influences: the external environment affects utilisation in ways beyond the influence of the primary stakeholders and the evaluation process. It includes indirectly involved stakeholders (not direct users) whose actions can affect the use (or non-use) of an evaluation. These include the public or media, governance structures (eg board, ministers, parliament), executive committee and donors.

These four areas are described in turn below.

Box 3.5 The RAPID Framework and the utilisation of the JEEAR evaluation

'How the Sphere Project Came into Being' (Buchanan-Smith, 2005) notes that the quality of the Joint Evaluation of Emergency Assistance to Rwanda (the JEEAR) was high and the findings actively promoted, targeted and disseminated. There were persuasive and key individuals championing its use, skilfully creating or exploiting alliances and networks. The evaluators themselves were skilled at negotiating the political landscape, identifying key individuals and groups to influence.

The evaluators also had strong links – common interests and close contacts – with the policy-makers. Many of the evaluators had direct humanitarian experience, enhancing their credibility and the legitimacy of the findings with the users. An important role was played by NGO umbrella organisations such as Interaction and ICVA. Specific individuals also had a powerful influence.

Finally, the external environment – the shock of Rwanda, the media spotlight – added to pre-existing concerns about performance in the sector and 'expanded the boundaries of what policy-makers would accept as necessary change'.

3.3.1 Quality factors

Of the four groups, factors related to evaluation quality receive the most attention in the literature on utilisation. This is not simply a question of the quality of the report, at least not in the usual sense: both the product and the process determine quality. How the evaluation is designed, planned, conducted and managed by client and evaluator matters just as much as the report, sometimes more.

Purpose and design

Section 3.2 above described instrumental, conceptual, process and legitimising types of use and the relative value placed on different uses according to the user's interests. According to Patton, the purpose, approach, methodology and presentation of an evaluation should derive from the 'intended use by intended users' (Patton, 1997). One size does not fit all, and 'each purpose privileges different users' (Williams et al, 2002 p35). 'Evaluations are situated in the context of various stakeholders' interests. They are by nature political' (Carlsson et al, 1999, p51). To optimise utilisation, the interests, needs, influence and power of the users should be identified at the outset and used to inform the design of the evaluation.

Patton's insistence on designing around 'specific, identifiable people, not vague, passive audiences' (Patton, 1997, p 382) presents an obvious challenge to evaluations intended for multiple potential users, such as interagency evaluations. He recommends a negotiated prioritisation of users and uses. Evaluations often include mixed objectives and multiple stakeholders, without prioritising or considering what this may mean in terms of approach. If an evaluation is intended to be used to improve an operational programme, then the primary users should be the operational staff. An analysis of their needs and interests would form the basis of the design which may, for example, emphasise beneficiary feedback and certain programme aspects above others. Ownership of the findings by the operational staff responsible for using the evaluation would, in this case, be essential; the written report, less so.

Alternatively an evaluation intended to inform organisational strategy would need to identify the stakeholders with the power and influence to effect changes at this level, and the evaluation should be designed with their needs in mind. The type of information and reduced level of programme detail that would serve their purpose would be of less use to operational staff.

Multiple purposes may displace each other by virtue of overload in the evaluation's scope, and one purpose may unintentionally undermine another. Accountability for one set of stakeholders may displace the intended learning use for others. For example, from the point of view of those whose work is being evaluated, the knowledge that judgements will be made and communicated in writing can create defensiveness. This is in direct opposition to the safe, trusting, non-judgemental atmosphere that modern approaches to adult learning strive hard to achieve.

At the same time, van de Putte's, 2001 study found that respondents differed in their experience of combined accountability and learning uses; some saw no contradiction. What appears to be essential is clarity and agreement about the purpose. The lack of this leads to 'everyone expecting that the evaluation would serve their individual information needs' (Williams et al, 2002, p20); some or all the users will be disappointed. Contradictory assumptions about use are also created by unexplained language. Terms such as learning, accountability and transparency are used loosely and often interchangeably.

Participation

Participation is fundamental to ensure that an evaluation's purpose and design are relevant to the users. The 'active participation of stakeholders at all stages in the evaluation cycle' (Williams et al, 2002, p53) promotes use. Participation is not passive listening or the act of being consulted. It demands meaningful and sustained involvement from those given the potential to make decisions about the evaluation process. 'High-quality participation is the goal, not high-quantity' (Patton, 1997, p383). Exposure to the process and learning also means that participants are 'in the best position to internalise these lessons' (OECD, 2001, p27) and more likely to value and trust the findings (Ridde and Sahibullah, 2005). Forms of participation depend on the context; they can include steering committees, regular group or individual discussions between the evaluators and users, workshops and advisory groups.

Participation throughout an evaluation also reduces the potential bombshell of a critical assessment or the 'shock' at recommendations for significant change. Individuals and organisations are more disposed to change if they are familiar with the information and mentally prepared. Studies often mention the low involvement of senior managers during an evaluation. Their involvement, if at all, occurs often only at the report stage. This may be significant for the low utility ascribed to evaluation by senior managers in several of the studies assessed. Sudden exposure

to proposed changes that are complex and demanding increases the risk of wholesale rejection or its equivalent – indefinite postponement.

Planning

Good planning also supports utilisation. Allowing adequate lead-time facilitates appropriate recruitment, wider consultation in the design stage and a good fit between the evaluation and key decision dates such as funding cycles and annual programme planning. A study of MSF(H)'s use of evaluations (van de Putte, 2001) notes that in several cases evaluations were not used because they took place too late: the decisions they should have influenced had already been made. Timeliness plays a particularly important part in humanitarian evaluation, particularly in field-based evaluations of short-term emergency programmes and RTEs when the window of opportunity is particularly brief.

Quality of the evidence

A great deal of emphasis has been placed on the content and form of evaluation reports. ALNAP annually assesses the quality of evaluation reports with respect to a set of criteria developed in its Quality Proforma. While studies of use have demonstrated that the (written) product is not sufficient in itself to guarantee utilisation, it is nonetheless an important factor. If the quality is poor, data dubious and recommendations irrelevant, it is perfectly rational not to utilise an evaluation. Although methodological rigour is important to ensure that collected evidence is credible, this is less of an issue for users than is the accessibility and relevance of the report. Time and again, excessive length and inaccessible language, particularly evaluation jargon, are cited as reasons for non-use.

Feasible, specific, targeted, constructive and relevant recommendations promote use. This is partly related to the user-focus and participation discussed above. Relevant, viable recommendations are more likely to be conceived if the users have been involved throughout the evaluation, communicating institutional and individual needs. Low relevance may explain why several studies noted that senior managers were the least likely to use evaluations. Although this is partly related to organisational and political issues, discussed below, it may also be similar to the experience of managers in the Finnish study who find that 'evaluations are often too backward-looking, they tend to say things with the benefit of hindsight. This kind of evaluation is seen as contributing very little to the management's work which essentially consists of finding future orientation' (p 11).

This resonates with the profile of senior managers described in the personality profiles such as Myers Brigg analysis as people who are future-focused and problem-solving by nature. Evaluations that do not provide some anchors for positive action because there is a critical rather than constructive problem-solving emphasis are less likely to be used

Some writers think that it is not the evaluator's job to make recommendations. An evaluator should generate the findings but it is the users' job to translate them into recommendations or actions. Anyway 'evaluators seldom have the expertise to make recommendations' (Patton, 1997, p327). Patton believes that the solution is the generation of recommendations as a collaborative process

Mechanisms for follow-up

How an evaluation is disseminated and followed up is another important utilisation factor. This entails far more than the distribution of the report. 'Many of our cases have shown that the evaluation report is over-emphasised, at the expense of other ways of disseminating findings' (Carlsson et al, 1999, p45). Dedicated follow-up individuals, the clear allocation of responsibility and specific mechanisms for action increase the likelihood of evaluation use, particularly if follow-up was planned from the beginning of the evaluation.

Examples of follow-up mechanisms include management action plans, follow-up working groups, evaluation communities and networks, the European Commission's *fiche contradictoire* (which invites a systematic response from potential users) and the creative use of the internet to disseminate findings and invite consultation. Participants at an OECD workshop in Tokyo (OECD, 2001) shared innovations such as CIDA's development of pop-up reminders for staff that appear at key planning and approval moments, linked to a learning database. The World Bank is investigating the utility of types of search engines based on linguistic analysis, capable of 'zeroing in on relevant material' (OECD, 2001, p 30). Other studies found that the incorporation of evaluation findings into training programmes enhanced use, although this is relatively rare and under-resourced.

An important consideration is matching the form of dissemination to the users. There is some evidence to suggest that adapting the mode of dissemination to suit the user is a stronger predictor of utilisation than designing the evaluation around their needs in the first place (Landry et al, 2001). 'Careful targeting allows intelligent tailoring of feedback approaches' (OECD, 2001, p 22). Danida for example has

identified three audience groups and uses different formats to suit different users, including a shortened 'popular version' of a report, newsletters, lectures and seminars. Users in some studies also requested reports that were customised for content, not just length, identifying the particular interests of, for example, senior managers and extracting relevant strategic issues. This, however, appears more likely to occur through verbal debriefs, and several of the OECD agencies emphasised the continuing value of face-to-face communication.

The evaluators

Although some of the studies (eg van de Putte, 2001) note that finding experienced evaluators is often a problem, surprisingly little is said about the competence of the evaluators. Their role with respect to utilisation is also notable by its absence. It is rare to find reference to the use of evaluators as a follow-up resource or trainers, or to any involvement of evaluators beyond the final report and workshop.

The evaluators themselves may not regard utilisation as their responsibility. The work predisposes consultants to an itinerant and potentially piecemeal approach. Patton departs from the traditional notion of evaluators as a type of researcher detached from the operational – and political – realities of utilisation. He argues that the independence and integrity of an evaluation is preserved through adherence to professional standards and guidelines, not through disengagement from the users' needs and context. He considers that evaluators have the responsibility to facilitate use, secure the attention of key users and understand how to make evaluation information compete successfully with other organisational imperatives. Further, evaluators should 'train users in evaluation processes and the uses of information' (p383).

To do this would require adequate time allocation for such quality interaction. Little is said in the literature about this, and the extent of such involvement in practice is difficult to ascertain from reports. For evaluators, assessing the potential for a utilisation approach before starting is easier said than done. Essential preconditions, such as the client's degree of motivation and availability to engage in the process, are hard to discern at the contractual stage. Proxy indicators, such as difficulty in obtaining access to key staff members for interviews, may quickly reveal the real engagement of stakeholders in the process, but by then it is too late; 'it's better to find out before preparing the meal that those invited to the banquet are not really hungry' (Patton, 1997, p85). If evaluators seek to ensure that their work is to be

used, far more pre-evaluation engagement (including participation in the development of the ToR) would be necessary.

Clients cannot judge the ability of an evaluator or team to facilitate utilisation, and the general absence of accreditation means that competencies are already varied and difficult to predict. The recognition of conceptual and process use implies a range of evaluation skills that exceed fluency with traditional evaluation methodology. 'A need exists to rethink the professional profile of evaluators that are to engage in enhancing learning. Next to research skills, such professionals should excel in communication, process facilitation, conflict resolution and negotiation skills. Analytical skills and tools will have to include situation and stakeholder analysis, as well as group dynamics' (Carlsson and Engel, 2002, p9).

The array of quality factors affecting use supports the importance attributed by the evaluation sector to good practice. At the same time, evaluation does not take place in a vacuum. Its use is also a function of the links between the evaluators and managers and the users.

Box 3.6 Summary of quality factors promoting utilisation

Design – purpose and approach The purpose is clarified and agreed among the key stakeholders at the design stage; the potential for multiple, potentially contradictory purposes is avoided through discussion and prioritisation. The approach is designed in accordance with the purpose and the users' needs and interests (eg a field- learning or programme-improvement purpose should ensure maximum involvement of the field staff in the design and a reduced emphasis on other stakeholders' needs such as for reports).

Participation and ownership There is meaningful, quality participation of all key stakeholders throughout the evaluation. Users directly influence the purpose and design, increasing its relevance and ownership, and are appropriately involved throughout the process (through clear mechanisms such as steering groups, discussion, clarification, regular feedback throughout the process).

Planning There is sufficient lead-time and adequate allowance for quality staff investment in the evaluation. The evaluation coincides with key decision-making

CONTINUED

Box 3.6 Summary of quality factors promoting utilisation *continued*

cycles or events such as a new programme cycle, new budget year/allocation or parliamentary debates. Unplanned timing can also be a factor – such as if the evaluation coincides with an unforeseen opportunity, event or crisis that increases its relevance. Lastly, timeliness is important. The utilisation window of opportunity is particularly short for operational humanitarian programmes and easily missed by relatively minor delays.

The evidence The evidence is credible (well researched, objective, expert). The report is easy to read (concise, with accessible language and no jargon). Recommendations are specific, prioritised, constructive, relevant, feasible and identify who is responsible for action and when.

Follow-up mechanisms Specific follow-up plans are established at the outset: clear allocation of individual responsibilities; a 'champion' or key person is committed to action; formal management response mechanisms and action plans; pro-active dissemination through promotion of the findings in user-relevant ways (eg through management meetings, presentations, seminars, team discussions, fast-track debriefs) as well as wide distribution through, for example, the internet. Use may be enhanced through customised dissemination, with content as well as form extracted or adapted for specific users (usually senior staff members).

Evaluator credibility The evaluator is credible (in terms of competence and reputation). The ability of the evaluator to be balanced and constructive is also important; wholesale negativity tends to lead to wholesale rejection. The evaluators and evaluation managers understand the political nature of evaluation, facilitate utilisation and manage stakeholders accordingly. A broadening of skills to include facilitation, stakeholder analysis, the management of conflict and group dynamics are important for user-orientated evaluation.

3.3.2 Relational factors

Efforts to ensure the independence of evaluators and evaluation units are essential to protect the credibility of the findings but can inadvertently undermine use. 'Some

agencies said that it is fundamental to the role of evaluation units, and is in fact their key attribute. Others argued that it can be a hindrance if independence becomes a barrier to partnership and dialogue' (OECD, 2001, p39). Independence can lead to a perception that evaluation is too far removed from operational and organisational realities. If evaluation is regarded as a somewhat academic exercise carried out by specialists, it can be viewed as opaque and resistant to users' needs. 'The motivation with regard to evaluation has been extinguished because everything in that respect comes in a ready-made package from the evaluation unit with the feeling that we have no influence whatsoever over it' (MFA Finland, 1997, p17).

Landry's study of research utilisation in Canadian public administration found that one of the best predictors of use is the intensity of the linkages between the researchers and policy-makers. The ability of individual evaluators to establish relationships based on mutual respect and trust also appears to be important. This is not to say that relations must necessarily be particularly friendly or informal. Cordial and professional relations between evaluators and clients are realistic expectations and appropriate for the job. At the same time, as noted by the Sphere case study (Buchanan-Smith, 2005), common ground between the evaluators and the policy-makers is important; the JEEAR evaluators came from the same 'community', establishing 'chains of legitimacy' (p102) and credibility.

Attitudes on both sides that generate a sense of two cultures or communities are unhelpful. Relational issues are sensitive topics and of course do not appear in the products of evaluations to permit review. This may be an important step to consider in the management of evaluations, and to help build trust in the process.

3.3.3 Organisational factors

Evaluation as a tool for organisational change

Individual learning contributes to organisational learning. 'But since organisations are not like individuals – and should not be anthropomorphised – they "learn" through processes that reflect organisational interests, structures, functions and decision-making context' (Suhrke 1999, p70). The complexity of organisational change means that it is hardly surprising that evaluation's contribution to the process is at best indirect. Recommendations aimed at policy or structural changes

Box 3.7 Summary of relational factors promoting utilisation

Personal and interpersonal The evaluator is able to establish constructive relationships with key users. Trust is established. The evaluator's interpersonal skills and commitment to quality can be important. Commonality of background and skills between evaluator and users enhances credibility. Overall, the perceived credibility of the evaluator is important.

Role and influence of an evaluation unit The evaluation unit manages to maintain its independence from decision-makers while ensuring close integration and relationships. The unit is able to play a mediating role between stakeholders. A key feature is to establish common ground and shared relevance between evaluation, policy and programmes through communication and relationship building.

Networks, communities of practice Key stakeholders (users or evaluators and evaluation units) have links with broader networks, influential fora and individuals, which enhances the credibility of the evaluation, extends its reach and ability to influence wider policy

in an organisation imply consensus and action from a sometimes-fantastic number of stakeholders. Several of the studies noted that findings that challenged strongly held beliefs and behaviour embedded in an organisation's culture were less likely to be implemented than technical fixes and discrete units of change. Utilisation is also higher when an evaluation dovetails with other initiatives for change. Even the highly influential JEEAR was most successful where it added momentum to existing initiatives.

A paper on the future of humanitarianism argues that organisations seek, above all, consistency and predictability, and shy away from complexity:

> The more a matter is seen to offer an immediate and direct solution to problems at hand, the more it catches the attention of those who determine policies and make decisions. Conversely, the more ambiguous and less

> immediate the information, the greater the divide between the strategist, the planner and the decision-maker. (Kent, 2004, p8)

This paper also notes that, faced with lessons learned that apply across the organisation, humanitarian agencies have a tendency to divide them into smaller, manageable chunks that can be packaged out to individuals and departments. This behaviour may partly explain the attraction of the matrix approach to evaluation follow-up. Though efficient and good for individual accountability, the division of labour swiftly removes the process of intra-organisational cooperation and dynamic debate that is needed for organisational change. The same is true of sector-wide learning.

Obstacles to organisational learning

Van Brabant's study (ALNAP, 1997) describes a number of other obstacles to organisational learning:

> centralisation and hierarchy, internal power structures and poor information management… having to live up to the false image that development and relief are quick and easy, the temptation to hype up one's performance in the face of growing competition between agencies, the financial instability of certainly *(sic)* humanitarian aid agencies, caused by a trend to go for cheap growth driven by short term funding and unrealistically low overheads, and the high degree of job insecurity of many staff. (p3)

Cautious leadership and 'the governance structure of a charitable board of trustees tends to exert a centralising and risk-averse influence. Board members (who in many countries cannot be paid for their work) may lack time or experience of development processes or of NGO management' (Mango, 2004, p4). Other impediments described in the literature include an organisational culture that does not value learning, staff members who do not understand evaluation, bureaucratic imperatives such as the pressure to spend regardless of quality, and the lack of real incentives to change.

The unequal nature of the aid relationship is also a significant barrier. Why and by whom an evaluation is commissioned affects ownership and hence utilisation. A paper prepared for an OECD/DAC workshop on evaluation feedback found that if an

evaluation was perceived, rightly or wrongly, as donor-driven, the impact was 'sharply reduced' (MFA Netherlands 2000, p5). The same can be said of evaluations viewed in the field as serving only headquarters needs, not the needs of the programme.

Challenges in linking policy with practice

High staff turnover and poor knowledge management are well-known impediments to embedding organisational learning in routine working practices. Furthermore 'evidence shows that NGOs are not tightly managed entities. Senior field staff work with considerable latitude, using their own judgement. Organisational policies are not always followed in practice' (Mango, 2004, p4).

Performance issues can also inhibit use. Just as utilisation is enhanced by motivated individuals willing to 'champion' the findings and promote use, it is also constrained by individuals who block or fail to act. There is some evidence to suggest that evaluations are sometimes used to compensate for the inability of management to persuade staff to adopt new policies and practice, with varying success.

Organisational imperatives

Evaluation reports can present a risk to an organisation's reputation. 'A lot of time and energy has been wasted in dealing with the public out-cries caused by the negative handling of an evaluation by the media', displacing the utility of the evaluation (MFA Finland, 1997, p11). The perceived risk may lead staff members to suppress and reject findings in the interests of protecting their survival. 'The need to maintain a high profile and operate within an increasingly competitive environment encourages a preoccupation with survival and self-justification across the sector, where organizations may lose sight of their higher ideals, focussing instead on activities that ensure their own continuation and growth' (Wigley, 2005, p23).

A study on obstacles to organisational learning notes a number of such 'paradoxes' in aid agencies. In Sida for example:

> Managers were supposed to spend the funds allocated to their programmes and to their countries, but too often it was impossible to spend the money fast if it was supposed to achieve the programme goals,

> whether they were reaching the poor or building sustainable institutions. They were forced to spend money on badly prepared or ill-conceived projects in order to avoid being criticised and seeing their budgets cut in coming years. (Edgren, 1999, p54)

There is also a limit to the absorptive capacity of the humanitarian system. It is often described as under considerable strain. Workload and the increasing number of upwards-accountability demands further reduce the space for learning. Time is a significant issue in the capacity of personnel to absorb learning and address change that requires sustained investment. Referring to understaffing and 'general working conditions approaching the unreasonable' (p18), the Finnish MFA study notes that as a result staff members concentrate on immediate demands; 'evaluation products never have the top priority because they do not require immediate action on the part of the staff member' (p18). The force and dominance of internal organisational imperatives imply that it is easier to discount performance lessons from evaluations than to ignore a threat to personal and organisational reputation.

It is clear that the majority of the characteristics and impediments to organisational change are beyond the control of evaluation. Facilitators and predictors of utilisation in general relate to the presence of positive structural and cultural characteristics that predispose organisational learning. In larger organisations, the existence of a well-resourced evaluation unit is, for example, an important determinant of use. The European Commission study concludes that Directorates-General with more institutionalised evaluation functions (a dedicated unit with significant evaluation experience, skills and established procedures) are more likely to use evaluations. Close structural links between evaluation managers and decision-makers are also significant.

3.3.4 External influences

External pressure may be a significant utilisation factor according to some studies. 'Many authors however feel that most change results from external pressures' (ALNAP, 1997, p3). The question is whether it is (published) evaluations that provoke external pressure. The ODI Sphere case study concludes that the humanitarian system is 'most responsive to change under pressure when the push factors are high' (Buchanan-Smith, 2005, p98), referring to the media and donor pressure

Box 3.8 Summary of organisational factors promoting utilisation

Culture Senior managers promote a culture of learning (openness to scrutiny and change, embedded learning mechanisms, transparency); staff members value evaluation and have some understanding of the process. Attention to performance is integral to working practice, managers actively support staff to learn and the organisation's leaders promote and reward learning.

Structure An evaluation unit or individual dedicated to accountability and learning is in place. The evaluation unit is structurally closely linked to senior decision-makers, adequately resourced and competent.

There are clear decision-making structures, mechanisms and lines of authority in place. Vertical and horizontal links between managers, operational staff and policy-makers enable dissemination and sharing of learning. There are permanent and opportunist mechanisms for facilitating organisation-wide involvement and learning.

Knowledge management The organisation has functional knowledge-management mechanisms, including systematic dissemination mechanisms, informal and formal knowledge-sharing networks and systems.

following the Rwanda crisis. Donors exerted pressure on UNHCR following the Kosovo evaluation, which, according to UNHCR's RTE Afghanistan, directly influenced its programme there (UNHCR, 2001, p14).

Yet there is little evidence in the studies of utilisation to suggest that evaluation is typically used by external stakeholders as a tool to effect change. As there is no regulatory body to monitor performance in the sector, evaluation remains an essentially internal reference without formal power. The recent Humanitarian Impact Conference, 2006 concluded that 'despite the accountability feature of impact evaluations, no consequences exist for organizations or individuals whose programs demonstrate poor performance' (Fritz Institute, 2006, p2). Despite the unusual influence of the JEEAR, its impact 'was also partial. The more challenging and radical recommendations around accreditation and regulation of NGOs were ducked and have been consistently evaded' (Buchanan-Smith, 2005, p102).

The issue of greater regulation and certification has been raised again by the TEC synthesis report, ten years later. The TEC synthesis also notes the absence of real external drivers for change. 'This lack of external pressure for change is one of the critical reasons why performance has improved so little in the last 10 years (including the performance of donors)' (Telford and Cosgrave, 2006, p109).

Protecting reputations and funding

This is not to say that the reaction of external stakeholders to published evaluations does not provoke a response from humanitarian agencies. The *fear* of repercussion from publicised criticism is certainly very real. However, agencies seem most likely to be galvanised to protect their reputations. The rush is not towards the recommendations. To complicate matters further, it does not appear to be performance that influences the funding decisions of either donors or the public. Other influences are at play.

A 2005 report by the UK's National Audit Office examines DFID's engagement with six multilateral agencies: 'Internationally, the effectiveness agenda is rising in prominence but is not yet at the forefront of funding decisions generally. Other donors surveyed felt that effectiveness was just one of a number of issues to consider when deciding who to fund' (NAO, 2005, p23). The ability to shop around is limited, in part, to particular global mandates and the availability of agencies on the ground. Moreover, donors are dependent on their partners and share a similar interest in painting a positive picture of performance. The increased presence of donors in the field has also reduced the power of evaluations as a source of information.

Many NGOs rely heavily on public donations. A loss of credibility through reports of poor performance would surely affect such voluntary funding. However, a recent survey of UK charities found that the public gave to charities based on trust and a belief that the money is spent 'wisely and effectively', despite a 'distinct lack of public scrutiny, and scant knowledge, over how charities are managed' (Charity Commission, 2005, p23). The most persistent public concern regarding charities continues to be the amount they spend on overheads, not their performance (Slim, 2002).

There is little in the studies to advise evaluation managers on how to manage the public's (unrealistic) expectations and the risks associated with transparency. Efforts to educate the public and media are an obvious route and one that some organisations such as the British Red Cross are pursuing. More honesty about the

complex reality of humanitarian work is a necessary but potentially risky step as agencies currently design their funding messages around the public's expectation of a simple and effective response administered by worthy outsiders. Addressing many of the performance weaknesses identified by evaluations (such as greater beneficiary participation) could alter the product–process ratio unfavourably and incur exactly the kind of increased overhead costs so mistrusted by the public.

There are moves to counter the impasse. According to the OECD workshop report, Danida's analysis is that damaging headlines are the result of attempts to conceal and the inevitable lure of an exclusive. In response Danida has increased the availability of evaluation reports and routinely issues press statements backed up by press conferences to permit intelligent debate. Japan has even included journalists in evaluation teams (OECD, 2001).

Pressure from the beneficiaries?

External pressure may come from the public or government of the recipient countries, seeking better accountability for the impact of money spent in the name of their population. 'Tolerance for the evaluation gap is waning. Developing country governments are demanding better information about the efficacy of social spending' (CGD, 2006, p3). This may yet become an influence on humanitarian funding, though the obstacles are higher and the likelihood low in countries with failed states and populations without voice.

The 'customers' should constitute an external pressure. Although beneficiaries are more likely now to be consulted during evaluations, it is not common practice to communicate the findings afterwards.

> Humanitarian agencies are perceived to be good at accounting to official donors, fairly good at accounting to private donors and host governments, and very weak in accounting to beneficiaries. This points to the very heart of the challenge confronted by those who wish to promote humanitarian accountability. The people whose welfare is meant to be the object of the exercise have the least say (indeed, often none at all) in designing policy or shaping operational practices. (HAPI, 2006, p3)

A study examining the motivation of governments to take development evaluation seriously (Gordillo and Anderrson, 2004) notes that: 'producing good quality

evaluations studies alone are not sufficient to motivate political action' (p13) 'The role of beneficiary populations in the evaluation design and implementation seems to be particularly important in creating incentives that support action' (p13) and 'downward accountability was crucial in motivating politicians to take the results of the studies seriously' (p13).

Of course there are numerous reasons why it may not be practical or even feasible to provide beneficiaries in affected communities with the results of evaluations. In some cases the political context means that findings may be too sensitive to share. At a community level, if the beneficiaries have not been active participants in the programme's design, then it is probably too late to engage at the evaluation stage. Mechanisms for feeding back to dispersed and vast beneficiary populations would have to be carefully planned and targeted; they may be impractical (due to problems of access) and could take as much time as the conduct of the evaluation itself. However, the marked absence of any intention to provide downwards feedback deserves further attention. There are examples of good practice such as Action Aid's new Accountability Learning and Planning System (Alps), which emphasises downwards accountability and participatory review. Such initiatives could be encouraged and provide a powerful moral incentive to act on evaluation results.

3.4 Utilisation in practice

This study sought to explore utilisation further through new case studies, a greater emphasis on the experience of the evaluators and particular users such as senior managers and through analysis of some sector-specific datasets such as an analysis of evaluation ToRs and a review of previous ALNAP work in this area. The types of use and factors affecting use described above provided a framework within which to analyse and test the findings. Four case studies were volunteered by ALNAP members for this chapter, specifically from CAFOD, MSF(H) and OCHA, plus a shared USAID–DFID evaluation (Box 3.9).

3.4.1 Types of use

MSF(H)'s use of an evaluation of its nutritional intervention in Bahr el Gazal (Valid International, 1999) is an example of the complexity associated with conceptual use. The Bahr el Gazal evaluation did not, by and large, lead to direct policy or practice changes, at least not at the time. Rather, it was initially rejected by many in the

Box 3.9 The case studies used for this chapter

The **MSF Holland c**ase study focused on the impact of three evaluations on the development of MSF's nutrition policy and approach: *MSF in Catastrophe* (Valid International, 1999), *MSF Holland in Afghanistan Mission Evaluation: May 2000 – May 2002* (O'Reilly and Shoham, 2002) and *MSF Holland's Medical and Nutritional Strategy in Darfur – Sudan* (Corbett and Lloyd, 2005). Six MSF personnel and the three evaluators of the 1999 and 2005 studies were interviewed for the case study.

In 2005, **CAFOD** commissioned *Rudder and Sails* (O'Donoghue, 2005) to 'Review the strategic priorities of CAFOD's Humanitarian function and provide clear feedback, analysis and recommendations on how CAFOD determines if, when and how to respond to a humanitarian crisis'. Six senior managers and the evaluator were interviewed for this study.

The **USAID Office of US Foreign Disaster Assistance (OFDA) and the UK Department for International Development (DFID) joint evaluation of the Humanitarian Information Centres (HICs)** was commissioned in 2004 to 'ascertain the effectiveness of HICs in servicing the humanitarian community' (Sida and Szpak, 2004). Five people were interviewed for this case study, from OFDA, DFID and OCHA, and including one of the two evaluators.

The Inter-agency Real-Time Evaluation of the Humanitarian Response to the Darfur Crisis 2005 was commissioned by the Emergency Relief Coordinator and Under-Secretary-General for Humanitarian Affairs (ERC/USG), Office for the Coordination of Humanitarian Affairs (OCHA) (OCHA, 2006). Eight interviews were carried out for this case study, with five OCHA personnel, one donor and the two external evaluators.

organisation. However, it generated new data, raised questions and proposed innovations that challenged MSF(H)'s thinking. It led to intensive debate (which continues) and to the development of alternative approaches in nutrition, linked to the ideas presented in the evaluation but adapted to MSF(H)'s culture and mandate.

The evaluation also influenced the evaluator. The inherent limitations of therapeutic feeding centres, particularly in contexts of overwhelming need such as Bahr el Gazal in 1999, led Valid International to develop an alternative approach (Community-Based Therapeutic Care, CTC) which has subsequently directly or indirectly influenced the nutritional approach of most major agencies in the sector.

The initially limited impact of the 1999 evaluation was due in part to the divergence of some strategic recommendations from MSF's organisational ethos and mandate. Several of the recommendations concerned community participation and an improved understanding of local cultures, a significant departure from MSF's medical approach to nutrition and its focus on centralised, expatriate-managed emergency response. The evaluation also predated the full development of a community-based approach; the ideas were in their infancy and difficult to operationalise. Particularly for senior management, the findings challenged (and could damage) MSF's reputation as a highly effective lead in the nutrition-response sector. The evaluation itself seeded a shift in thinking and contributed to a slow process of review and further development. Its 'use' was conceptual, rather than direct and instrumental.

CAFOD's review of its humanitarian function (O'Donoghue, 2005) stimulated new thinking and learning. The review took place within a broader process of organisational change, and has led to significant policy changes. In this sense, the use was instrumental, with action clearly linked to the evaluation findings. However, the recommendations themselves were the product of a highly interactive and iterative process which stimulated, 'downloaded and captured the ideas and thinking of individuals'. In other words, the recommendations did not spark the changes, they captured the dialogue facilitated by the evaluation and evolved from a process of change already in motion.

Through the process of the CAFOD review, staff became aware that they had not really 'understood' the new organisational policy. They 'knew' about it, and had signed up to it, but they had not understood what it meant in terms of operational practice, the capacity, roles and responsibilities of their departments. The review

clarified the implications of the new policy and of a changing context for CAFOD. This process learning also reinforced and developed learning derived from other sources. The cumulative effect was viewed as extremely important as it increased the credibility and force of the review's findings and contributed to longer-term conceptual learning in the organisation.

In 2004, USAID/OFDA and DFID jointly commissioned an evaluation to 'ascertain the effectiveness of HICs in servicing the humanitarian community'. Although there was some instrumental use, USAID, DFID and the Humanitarian Information Centre (HIC) manager commented that there was little in the evaluation that was new to them. The donor's involvement and the ongoing availability of mission reports and reviews meant that the evaluation contained few surprises. However, although DFID would have appreciated more radical recommendations, neither donor viewed the predictability of the findings as a weakness. It led to some instrumental use, but the most important use of the evaluation was by the donors because it substantiated and consolidated existing information (making it easier to share). Importantly, it confirmed and independently validated their perception of the HICs. It also provided a kind of benchmark and reference document about the HICs.

Similarly, MSF(H)'s 2005 Darfur evaluation provided confirmation of what was known or suspected; it legitimised a position. It provided an independent reference for ongoing discussions about the effectiveness of MSF(H)'s performance in nutrition in Darfur and elsewhere.

The case studies also demonstrate non-use. Although USAID and DFID both responded to some of the recommendations, there was little direct impact on OCHA. Some 18 months after the HIC evaluation, the new manager initiated a joint USAID, DFID, ECHO and OCHA review of the new Pakistan HIC, which noted that '[l]essons identified in both the 2004 USAID/DFID evaluation of HICs and the 2005 OCHA/ECHO monitoring mission to Sumatra and Sri Lanka have been only minimally implemented' (p3).

The case studies demonstrate the direct and indirect effects of some evaluations on organisational and individual learning. While some instrumental use is evident, the case studies reinforce the findings derived from the literature overall; use of evaluations is varied, sometimes unpredictable and with particular actions or changes difficult to attribute to a particular evaluation. Conceptual and process use is common, as is legitimising use. Instrumental use is the least likely form of

utilisation, and one evaluation can lead to a mixture of uses, as well as partial or selective use. Commissioning evaluations for purely ritual reasons, such as to meet administrative or legal directives or to symbolise organisational accountability without any real commitment to using the results, may be quite widespread.

3.4.2 Quality factors affecting use

To what extent do quality, relational and organisational factors explain the use and non-use of evaluations in the four case studies? What is the working experience of evaluators and evaluation managers, and what do the ToRs in circulation tell us about the way in which recent evaluations have considered use-promoting factors? The following three subsections (3.4.2 to 3.4.4) address these questions.

Design: purpose and approach

A review for this chapter of 30 evaluation ToRs, randomly selected from those distributed through the ALNAP network, shows that a combination of different purposes is common. Of the 30 ToRs reviewed, 22 contained two purposes, most often a combination of accountability (judging effectiveness) and organisational learning. Over a third contained a learning (or programme-improvement) purpose, but rarely in isolation. There was little evidence of variations in the evaluation design according to these different objectives.

Several of those interviewed for this study thought that accountability and learning purposes should be separate. If accountability is sought, then an audit approach would be more appropriate. The intention would be to measure performance against standards and indicators (such as Sphere, organisational mandates or international agreements and policy instruments). The audit would be linked to mandatory cooperation from the 'subjects' of the exercise, rather than through the semi-negotiated access common to evaluation. Learning objectives would become a separate, internal process designed around specific field needs. This would imply outcomes that do not rely on – or sometimes even produce – a written report. Interestingly, several evaluators mentioned the difference between what is said (and agreed) verbally with staff and the shift in perception (and defensiveness) when the same thing is put in writing.

At the very least, a mixed agenda should be explicit. Some of the ToRs reviewed for this study demonstrated a degree of coyness about straightforward accountability objectives; the purpose is often referred to as 'learning', without reference to the principal motivation of the evaluation, which is to judge and report back. This may not be intentionally disingenuous, but the lack of clarity (and courage) about the purpose can generate suspicion towards the evaluation and confusion for the users.

The sector is increasingly experimenting with new approaches to evaluation suited to particular uses. After-action review emphasises internal learning. Joint evaluations can address the specific issues of coordination and collaboration that single-agency reviews do not. Real-time evaluation (RTE) can focus on immediate feedback and learning-in-action that ex-post evaluation cannot.

It is too early to assess whether innovations such as joint and real-time evaluation are being used for their intended purpose or not. The experience of some joint evaluations described in Box 3.10 is mixed. Some of the advantages of joint evaluations have been realised, such as presenting an unbiased common picture of the whole response. Equally, the disadvantages which would be predicted by Patton's user-based approach are evident. Findings are too general, ownership is lower and the evaluation lacks specific stakeholders responsible for subsequent action. Many of these joint evaluations combine a review of interagency process with the sum of fully fledged individual evaluations of each agency. This may lead to an evaluation that is overstretched and misses all its targets, displacing its natural focus on how the system is working as a whole.

The case studies similarly demonstrated intermingled objectives. For example the Darfur real-time, interagency evaluation combined the objectives of immediate improvements to field performance and organisational learning for future humanitarian action. This combination of field and organisational learning objectives meant that the evaluation design had to fulfil immediate learning needs (eg a focus on the field, rapid feedback, an iterative approach) while also providing the type of information needed by senior policy-makers and managers (eg comprehensive reports). Additionally, the public availability of the report and the fact that it was viewed as driven by donor criticism of the response led to a perception in the field that upwards accountability was the main real purpose of the evaluation.

Box 3.10 Joint evaluations

The concept of joint evaluation has been promoted by OECD/DAC since 1991 but the experience has been patchy. A review in 1998 noted a number of limitations, including 'too general and diplomatic conclusions as they have to combine different interests'. It found that there was little enthusiasm for joint evaluations but, encouraged by the GHD initiative, donors have returned to the issue recently. There has been a DAC Evaluation Network Working Paper, a conference in Nairobi and publication of OECD/DAC 'Guidance for Managing Joint Evaluations', updating an earlier paper from 2000.

A review of the Niger evaluation by the participating agencies (CARE, Catholic Relief Services, Save the Children and World Vision) (Wilding et al 2005) concluded that the advantages of this joint evaluation were:

- presenting the 'big picture'
- lack of bias
- provision of a basis for setting up Niger's NGO coordination forum, and provision of direction for recovery activities
- shared costs
- a common picture for donors.

Among the disadvantages in this specific case were:

- lack of a focal point to take forward the evaluation recommendations
- agencies did not participate equally
- different approaches of the agencies, causing conflict within the evaluation team.

Overall, the NGO experience has been mixed. Another joint evaluation supported by the ECB, in Guatemala, was considered less successful by the commissioning agencies. The joint Tsunami evaluations by CARE, Oxfam and World Vision were also considered to be difficult experiences though they did assist each agency to improve specific programme work. The general lesson so far seems to be that joint evaluation is a complicated exercise requiring a level of engagement that stakeholders on the ground are not usually able or willing to provide.

The specific objectives in the Darfur ToR are wide-ranging and include most of the evaluation criteria that may be expected in ex-post evaluations. The overload evident in the ToR was in large part a hazard of a multiple-agency evaluation. The evaluation team also tried, appropriately, to respond to emerging needs in the field. As a result, the evaluation addressed issues as diverse as water quality and the role of the Security Council. Field staff in Darfur viewed the visits as 'dipping in and out' without being able to develop and deepen the analysis. There was little evidence in the case studies of any clear stakeholder analysis, either to identify specific users' level of interest and influence or to target the recommendations.

The accountability or organisational-knowledge purposes of the HIC and MSF(H) Darfur evaluations lent themselves to use at the organisational rather than field level. Although this meant that, for example, the MSF(H) field staff (rightly) viewed the evaluation as HQ-driven and (perhaps wrongly) as of little value to themselves, at least the purpose was clear. The approach (external evaluators, the buy-in of headquarters rather than field staff) largely matched the intended use of the evaluations.

According to the questionnaire and interviews for this chapter, it is relatively rare to design an evaluation with flexibility built into the ToR and budget, permitting adaptation in response to the context and emerging needs. While evaluators may voluntarily extend the time allocated to their work, this is more likely to occur before or after field trips. Matching the availability of multiple stakeholders with that of external consultants is extremely difficult. The interagency evaluation went to considerable lengths to be flexible but was constrained by the availability of staff in the field. Greater flexibility in the field would have to be planned in advance and built into the design and budget; for large-scale, real-time evaluations, this implies almost full-time availability.

Participation and ownership

The hypothesis of quality participation is that it will increase ownership and relevance. This is particularly important if the purpose is to generate learning. As noted in Section 3.3, crossing departmental boundaries to generate organisation-wide learning is vital but difficult and unusual. CAFOD's approach (Box 3.11) sought to do just this.

For Darfur, MSF(H) developed the ToR in Amsterdam, following a common process of then sharing it with the field and responding to suggestions. According to the evaluators, ownership was low in the field. There was no contact between the evaluators and the field before or after the evaluation. Had there been, despite the fact that the evaluation was delayed and the programme completed, there may have been opportunities to include objectives valued by the field staff (such as to contribute to new programme planning).

The interagency real-time evaluation (RTE) in Darfur was commissioned by the ERC/USG, and managed by OCHA. It required rapid deployment in order to be real-time. As a result the ERC/USG requested that OCHA HQ draft the ToR, which were circulated to IASC principals and then finalised without real field participation.

Box 3.11 Generating organisation-wide ownership

The organisation-wide ambitions of the CAFOD review (O'Donoghue, 2005) meant that the product had to be informed and owned by all of the senior managers. Established intra-organisational teams already in place provided an appropriate structure for organisation-wide engagement. The process adopted was highly participatory and iterative; a series of discussions took place and the ToR were adapted to evolving needs. The evaluator facilitated a series of group discussions, rather than concentrating on one-to-one interviews.

The effect of this approach was to capture the team and organisational dynamics, essential to a review pitched at an organisational level. This process of sometimes-intense debate meant that the recommendations evolved through an interactive series of edits, distillations and revisions. While the evaluator proposed options, the process itself meant that staff was engaged, 'translating' the findings, making them relevant to CAFOD and, most of all, owned.

The field viewed the evaluation as a donor-driven accountability exercise. Coming at a time when the response was scaling up in an intensely difficult context, the evaluation team was not, by and large, welcome. To make matters worse, agencies were suffering from a major shortfall in personnel; they 'needed staff, not scrutiny'.

This experience demonstrates the real challenge to an RTE and in particular to interagency RTEs. Rapid initiation is essential, but so is the full engagement of field personnel, given that they are the primary users. Had the ToR been designed in the field, the number of stakeholders involved and the complexity of achieving consensus would have severely delayed implementation. In fact it may not have happened at all; over-stretched personnel may not have elected to conduct an evaluation at that time.

A tentative conclusion is that an interagency RTE is an incompatible composite of evaluation types. In the future, a careful analysis of the trade-offs between timeliness and participation, as well as a ruthless prioritisation of purpose should underpin the design. The result may be a departure from the approach expected by some stakeholders and 'normal' evaluation practice. The first visit, for example, could be entirely dedicated to designing the evaluation, through a collaborative process involving the key field stakeholders and the evaluators. The result – in theory – would better suit the field users, but may not suit the information and accountability needs of others.

Planning

Most of the evaluators interviewed said that lead times for evaluations were often 'ludicrous' with far too little time to develop the process (an observation corroborated by the number of evaluations advertised through ALNAP which are expected to start almost immediately). Stories were told of last- minute contracts, ToRs being read on the plane, minimal contact with staff and few opportunities to establish and build relationships.

Similarly, many said they spend more time on an evaluation than had been allocated in the budget, indicating a mismatch between the planning and the reality – a perception not shared by the evaluation managers, who were more likely to think the time allocated was adequate. It may be that the managers are unaware of the time constraints experienced by evaluators. Either way, tight schedules reduce flexibility and ad hoc opportunities for learning.

An RTE also faces the particular challenge of taking place during particularly intensive periods of an emergency. The interagency evaluation team arrived in Darfur when agencies were intensively scaling up their activities. OCHA had also had 175 visiting delegations over the preceding 18 months.

Timing was an issue for both of the Darfur case studies. The MSF evaluation was delayed and took place too late to influence the programme, which was then approaching closure.

Last-minute planning, difficulties in recruiting evaluators and failing to allocate sufficient time may be more significant for low utilisation than is recognised. There is little recorded about such administrative constraints, and perhaps more attention should be paid to their impact.

The evidence

How the report is written is, according to interview findings, an important factor in take-up. The kind of social-sciences language sometimes found in evaluations is 'ghastly' and apparently reduces the readability of the report. Simple, concise and clear language was preferred. Aspects of language can be organisation-specific. MSF(H) for example dislikes academic or 'woolly' language which is ill-suited to MSF's straight-talking culture.

Most of all, interview comments related to the utility of the reports were linked to the length of the reports (too long and therefore not read) and the nature of the recommendations. Information overload is an increasing problem, and many personnel interviewed viewed the length of reports as an inhibiting factor. Recommendations are said to be too numerous, forming 'shopping lists' that are not prioritised or linked to identifiable individuals or even departments. Broad, strategic recommendations are difficult to action, involve multiple stakeholders and accountabilities and are less likely to be acted upon.

The extent to which recommendations should be generated through a collaborative process to ensure their relevance and feasibility is an interesting consideration for utilisation. This approach does not appear to be the norm. Interviewees suggested that it was common for the evaluation users simply to correct factual errors in a report. This is intended to preserve independence but ignores the fact that recommendations represent a process of operationalising the findings. Some agencies point out that external evaluators do not know enough about the organisation to do this. Others involved the evaluators as facilitators or combined internal and external evaluators. One of the case studies periodically referred back to the evaluator to ensure that the findings and spirit of the recommendations had

not been distorted or co-opted in the process, an interesting example of maintaining a balance between the independence and integrity of the evaluation and relevance to the organisation.

Follow-up mechanisms

Three of the four case studies used a matrix approach, with recommendations presented alongside actions, individuals or departments responsible, sometimes with a timeframe. One agency designed a clear mechanism for follow-up that was based around an individual focal point and was highly interactive. The focal point maintained momentum and played an active dissemination role – including the anticipation of concerns, additional information needs and potential difficulties before they could lead to blockages.

There was little evidence in our sample to demonstrate whether matrices and management plans of action promoted use or not. They certainly capture the findings and may be the only practical solution to following up on large evaluations. They can however become unwieldy. It is impractical to list individual responsibilities or posts if there are scores of recommendations and stakeholders. They cannot capture the complexities of implementing certain recommendations. Neither can they substitute for active management and accountability at all levels in an organisation.

> The ERC, OCHA and the IASC did not manage to get the full cooperation of the field for the exercise, but also did not insist on a more rigorous response and action from the field on the evaluation team's findings... Follow-up was constrained by several real and practical challenges, but it was also made possible by the lack of follow-up and ownership at the senior headquarters' level'. (OCHA, 2006, p3)

If action plans are not combined with a face-to-face approach, they can also invite an unproductive series of claim and counter-claim in which the user refutes a recommendation, the evaluation team responds, the user counters and so on. OCHA, initially relying on a matrix approach, subsequently recruited an experienced M&E specialist in a full-time follow-up post, based in Sudan. Follow-up took on a more strategic focus on systemic issues and worked on translating the recommendations that were still relevant into a dynamic management tool. OCHA

HQ personnel consider that this is an essential role; future evaluations will aim to have such a focal point in place earlier, facilitating and acting as a 'champion' for follow-up.

The sector is still learning how to establish meaningful and dynamic mechanisms for follow-up. The low involvement of the evaluators in this phase is striking; it may be pragmatic but surely undermines their ability to develop their craft (Box 3.12). The survey questionnaire showed that half the evaluators' clients rarely or never established follow-up mechanisms, giving an overall picture of weak or, at best, patchy follow-up planning. There is also little evidence of adapted and creative dissemination methods designed to suit different stakeholders; the norm continues to be a standard report and executive summary. This may indicate that the sector still has much work to do before the routine inclusion of follow-up mechanisms in evaluation planning is common practice.

The evaluators

The competence of the evaluators was not generally raised as an issue in the case studies. Credibility partly relates to the users. For example the studies indicate that field staff are more likely to view an evaluator as credible if they have an operational background and know the context (and preferably the organisation) well. A good skill match, such as skills in the management of organisational change for an evaluation leading to organisational learning, is important.

Box 3.12 Evaluators and follow-up

The lack of the evaluators' involvement following evaluations raises issues about their learning. Survey responses indicated that the provision of systematic and detailed feedback on the quality of the evaluators' work was patchy. None of the evaluators interviewed for the case studies had any idea of what had been done with their work or had any detailed feedback on the quality. Given the concerns about the quality of recommendations and the accessibility of formats and language, this is a missed opportunity for evaluators to increase their skills and the utilisation of their work. More involvement and feedback would also increase the ownership and accountability of the evaluators for the utility of their work.

Summing up quality factors

Evaluations that identify, prioritise and focus on the main purpose and primary users, designing the approach and product with their needs in mind, are more likely to be used. Planning for follow-up and making adequate time for meaningful participation and timely evaluation are essential to ensure relevance, ownership and systematic consideration of findings.

However, findings from research carried out for this chapter nonetheless suggest that the application of such pro-use approaches is patchy at best. In addition to relational and organisational factors, it is possible, though not at all proven, that this is due to limited awareness of quality factors that can enhance utilisation. There is quite a body of information available, but it is dispersed among numerous individual agency studies and research papers that many practitioners may not have seen. One key reference, *Utilization-Focused Evaluation* (Patton, 1997) runs to 400 pages of close print; not all will or can find the time to read – and then adapt – such information. There could be value in summarising the practices in a brief guideline or checklist format that is more accessible and easier to operationalise.

3.4.3 Relational factors affecting use

The second main group of factors affecting utilisation is based around links and relationships between the evaluation community and the users. MSF(H)'s nutritional policy was influenced by participation in networks and by its peers in the MSF Movement, themselves influenced by the community-based approach seeded in the Bahr el Gazal evaluation. In a sense, MSF(H)'s participation in sectoral and peer networks generated a boomerang effect, in which the influence of the evaluation returned to it via the networks and peers.

The facilitation and interpersonal skills of evaluators are an important influence. The ability of evaluators to create a learning atmosphere is important, if that is the purpose of the evaluation. It can involve skilfully managing the border between what remains unsaid and what becomes black and white in the report. The attitude of the organisation and its staff to learning and accountability will also affect the ability of an evaluator to exploit opportunities for active learning. In one of the case studies, the evaluator had worked for the agency previously; he was known and trusted. In other cases, the relationships were variable. On occasion, the frustration of field staff

with an evaluation was transferred to the evaluators. The personalities involved can be a significant factor, limiting opportunities for the evaluators to build relationships and undermining use from the outset.

There is a tendency to expect staff members to accept criticism gracefully and constructively. In reality, this often does not happen. The interpersonal and conflict-management skills needed to manage defensiveness and opposition are rarely mentioned as essential competencies. In general, the interpersonal dimension of the evaluation process goes unrecorded, despite the fact that it may be a significant factor in the quality and utilisation of an evaluation.

It is worth noting that evaluators were regarded with considerable suspicion by some of the senior and field personnel interviewed. The evaluator may be regarded as a harbinger of doom, or a nuisance and diversion from work. A common perception is that evaluators are too concerned with their reputations and less concerned about the utility of their work. Interviewees complained that evaluators did not recognise the provenance of many of the solutions proposed, which may have been identified by staff or already been implemented. Evaluators are sometimes viewed as unwilling to make information more user-friendly (by avoiding evaluation jargon) on the basis that their reputations would be damaged if they did not comply with the expected standards of a good evaluation report.

The commercial nature of the relationship with external evaluators can also be an issue, particularly in smaller organisations and for NGO field personnel who are more likely to regard consultancy fees as excessive and the work as mercenary. Some evaluators interviewed noted a tendency of clients or field staff to foist an almost inhuman schedule on the team, as if the most had to be obtained from every invoiced day.

Overall, insufficient attention is paid to the relational side of evaluation. Using evaluation is as much a people issue as it is a technical one – and perhaps more so. The findings of this study suggest that evaluators and managers need to invest more to generate trust and bridge the gap between users and evaluators. If the focus of the evaluation is on learning, then the evaluators and users should be in partnership, pooling experience and skills in the service of improved performance.

3.4.4 Organisational factors affecting use

Examining the learning cultures, structures and knowledge management of the case-study organisations was beyond the scope of this study. Of interest is the effect of management and political issues on their utilisation of the evaluations. None of the agencies involved, whether they have relatively hierarchical or 'flat' organisational structures, can simply communicate directives to change policy or practice and expect them to be implemented. Often, whether or not policy is changed in response to evaluations is dependent on the convictions of individual stakeholders. Decentralised agencies have highly autonomous field decision-making and the links between the programme and the headquarters sectoral or policy staff are relatively weak; headquarters personnel are largely advisory. Translating policy into action in the field can be difficult.

The relationship between the evaluation and decision-makers for the interagency Darfur evaluation could hardly be more complex. The ERC/USG commissioned the evaluation but does not 'line manage' the heads of the UN agencies in Khartoum. The Humanitarian Coordinator's authority over agencies in the field can also be contested, particularly by the NGOs. An interagency Core Learning Group (CLG) was established at HQ level, but was not populated by decision-makers. The Steering Committee for Humanitarian Response, representing the NGOs on the CLG, had no authority over the operational decision-making of its members in Darfur. The lessons-learned paper prepared by OCHA (2006) notes that 'a great deal of the lack of ownership can be attributed to institutional insularity, which led to a lack of accountability for the exercise at the individual agency, as well as collective IASC, level' (p4).

Concerns for an organisation's reputation featured in some of the case studies. In Darfur, agencies were concerned that 'highlighting and publicizing gaps in the response would dash donor confidence and thus discourage support at a time when low funding was seen as a key obstacle to improving the response' (OCHA, 2006, p3).

The sense of an evaluation being imposed was a clear issue, although in these examples it was usually an internal issue between headquarters and the field, rather than a reflection of unequal aid relationships between organisations. The electronic

questionnaire conducted for this study asked respondents what, in their view, was the most common reason for commissioning an evaluation, choosing between four categories of formal accountability (it has to be done, rarely commissioned by the organisation), improving the programme, organisational learning and legitimising. For the majority of the evaluators, formal accountability was perceived to be the primary motivation for commissioning an evaluation. This may indicate that 'ritual' evaluations are common.

The top-down imposition of an evaluation was cited as one of the three main factors inhibiting use by half the evaluation managers and over a third of the evaluators responding to the survey. If ritual and imposed evaluations are indeed common, then the potential for utilisation is greatly reduced even before the evaluation has begun.

MSF(H) and CAFOD have both taken proactive and robust approaches with donor-commissioned evaluations to ensure that the evaluation was of value to both partners. CAFOD and MSF are unusual in that a high percentage of their income is private. Other agencies appear less confident in regard to influencing the purpose, approach and use of evaluations they have not themselves commissioned. Good practice by some donors is in evidence through the delegation of evaluation to their funded partners, an approach largely adopted by USAID. ECHO is in the process of developing an evaluation methodology that can be managed and implemented by the partner NGO itself. The 'push' is still there, but the partner can design and manage the evaluation to optimise learning.

Stakeholders, including the evaluators, have the right to question intended use. The fear of losing funding or of appearing to be resistant to learning makes agencies timid about challenging the purpose of an evaluation and demanding greater partnership in the process. Similarly, all stakeholders have a duty to consider whether the financial and opportunity costs of evaluation can justify a symbolic or administrative need. Something cheaper and quicker may suffice.

3.4.5 External factors affecting use

This study has found little evidence of external pressure acting as a force for the utilisation of evaluations. In fact, it is noticeable by its absence. The questionnaire

responses showed that it was unusual for the key findings of an evaluation to be communicated to the programme's beneficiaries in the field. No cases were found in which an NGO, for example, held public feedback sessions with community leaders.

Most of those interviewed said that evaluation was important, but was only one of the resources and influences for change. It was generally given a middle ranking in terms of its value to decision-makers. Influential factors appear to vary, in part according to the particular focus or culture of the organisation. Some are strongly influenced by peers, others by technical reviews, visiting experts and participation in networks related to particular sectors. Senior managers often referred to the importance of discussion with staff. The use of a variety of sources of learning is of course positive. But it implies that evaluation planners and evaluators could do more to recognise the relationship between an evaluation and other learning processes. Playing a modest role does not undermine the value of an evaluation. There are other factors that do.

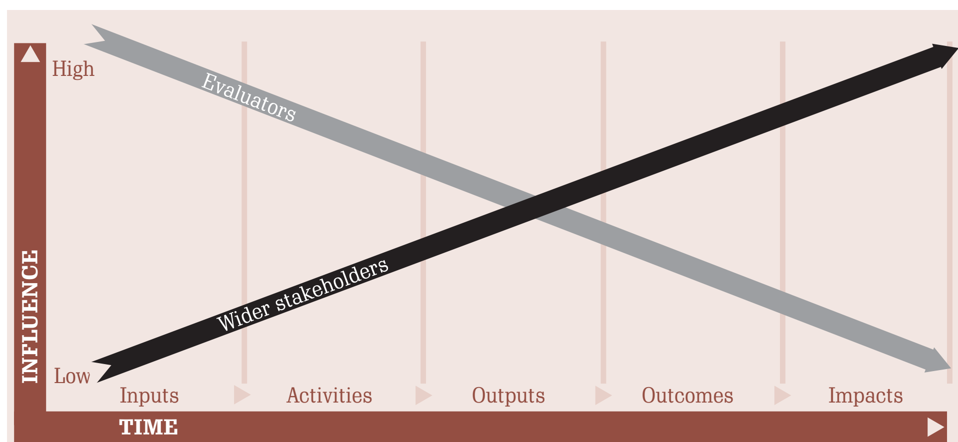
Most of the literature looks at change in only one direction – from the evaluation forwards to its utilisation. Although questions were included in this study regarding other influences on learning, the scope precludes any conclusion. Responses however indicate that most people and organisations do not seem to have given this a great deal of thought. What has worked? What are the most powerful agents of change in the history of an organisation, or in the sector? A rich line of enquiry could be to trace backwards from identifiable changes and explore the influences and pressures that led to it. A better understanding of evaluation's role in this process could enhance its design, or perhaps lead to the invention of different tools altogether.

3.4.6 Limitations to the utilisation of evaluation

Utilisation is only partly controlled by evaluation quality and process. Its ability to effect change in the context of powerful organisational obstacles and political imperatives is relatively modest. There are many competitors for an organisation's attention, and evaluation often loses out to other priorities. Evaluation is one (competing) part of a decision-making process. 'The policy-making process is a political process, with the basic aim of reconciling interests in order to negotiate a consensus, not of implementing logic and truth' (Weiss, 1977, p533).

The entirety of the utilisation process is not under the control of any one person or process. The more complex the use, the more this is so. Arguably, the evaluation is but a catalyst, influencing the beginning of a process and increasingly ceding control to the impact of other actors and other forces. Humanitarian workers have the responsibility to make sure that the inputs, activities and outputs of a project are the best possible basis for the intended outcome and impact of the programme, but they cannot predict exactly what those higher-level results will be. The influence of an evaluation similarly reduces in inverse proportion to its appropriation by the users (Figure 3.2).

Figure 3.1 The decreasing influence of evaluators in the change process



(Adapted from Ramalingam, 2006)

The core responsibility of the evaluators and managers is therefore to provide a good foundation for use. After that, the use is in the hands of the users – and so it should be. Although there may be an argument for the evaluators subsequently to play an advisory role (exemplified in the follow-up to the JEEAR), neither evaluator nor evaluation can be the sole agent of change. In addition, there are limits to the change that the sector can bring about alone.

Some of the angst surrounding non-utilisation may be misplaced. In part, it is the 'recurring structural impediments' (ALNAP, 2005, p25) that lead to the conclusion that evaluation is failing to contribute to improved performance. In part, this is true. At the same time, many persistent weaknesses relate to structural characteristics of the humanitarian sector that aid agencies argue are beyond their control or mandate.

The main recurring themes identified each year by the ALNAP synthesis are: human-resource problems (such as lack of training and trained staff, high turnover); competitiveness and poor coordination; weak links between relief, rehabilitation and development; the politicisation of aid and capacity building of national partners. Many of these weaknesses are at least partly the consequence of unpredictable funding and the earmarking of income for operational response rather than institutional development.

As long as aid is linked to government money and the public's expectations of charity in the north – which still expects white faces, quick solutions and good visuals to illustrate good performance – there are limits to the sector's latitude for change. Evaluation recommendations that, for example, propose reduced staff turnover and longer contracts may not be implemented because it is simply impossible for the agency to find, retain and fund those personnel.

Conversely, structural issues can be used as an excuse for inaction and conceal an unwillingness to challenge the status quo and seek innovative solutions. Should an evaluation concentrate only on what is within the control of an agency to change or insist on reference to an aspirational yardstick? Comparison with an ideal performance will find all humanitarian response wanting. There is an argument for evaluators at least to differentiate between what the potential users could do, and what should be done if other conditions were met. Evaluation is then fully contextualised at all levels.

This would again increase the need to target the relevant decision-makers. The decision-makers who can influence the systemic issues undermining performance are not typically the audiences of evaluations. USAID may seek to increase funding for Darfur but it is not the US Congress. Evaluation appears, at any rate, to be poorly suited to influencing political agendas. How much effort is made to ensure that the system-wide problems are 'translated' and communicated for the consumption of the right audience, or are we just talking to ourselves?

At present, evaluation is expected to serve both the day-to-day and loftier objectives of the sector. As well as serving learning needs, administrative formalities and information provision, there seems to be an expectation that evaluation will also generate organisational restructuring, strategic reorientation and influence on the wider political landscape in which the sector operates. This suggests either that evaluation as a tool has to adapt considerably to suit these different objectives, or

that broad recognition is needed that evaluation makes a valuable if modest contribution to a menu of tools and strategies for change.

Evaluation viewed as one of several tools at once enhances its potential to influence and demotes it from its position as a catch-all for accountability and learning. As other studies have noted (eg Ramalingam, 2005), learning strategies work best when they are an integrated component of a dynamic, phased and progressive approach to organisational development. Perhaps this, rather than the isolated utility of evaluation, should be the primary concern and focus of future work on learning and accountability.

3.5 Conclusion and ways forward

3.5.1 Concluding remarks

This chapter began with a concern that the credibility of evaluation will be undermined if its poor record of influencing humanitarian performance continues. In some ways then, the findings indicate a less gloomy outlook: our pessimism partly results from a narrow perception of utilisation that does not do justice to the rich and often indirect use and influence of evaluation.

At the same time, this picture of utilisation makes it difficult to assess the effectiveness of evaluation itself. Its utilisation is largely a mirror of the complex nature of change; it is transformed by the users themselves, strongly affected by context and often unpredictable. Evaluation is also irrevocably bound up with the learning culture and absorption capacity of the organisations and individuals involved. 'Serious attention to use involves financial and time costs that are far from trivial' (Patton, 1997, p383). A user-based approach is also challenged by the sheer number and geographic distribution of stakeholders that characterise humanitarian response. This means that even if the evaluation community gets better at applying a utilisation-focused approach (as it clearly should), there are real limits to how often such evaluations should be conducted and to the changes that evaluation can be

expected to lead to. This is true of any single learning and accountability mechanism that must also compete for attention with a diverse and often political range of organisational pressures and individual learning preferences.

If we continue to expect evaluation to cover most of the accountability and learning needs of the sector, we will be disappointed. It is critical that the job expected of evaluation is realistic, clear, better differentiated from other objectives and designed accordingly. Evaluation is often squeezed into multiple roles for which it is not always best suited, or that contradict each other. It may not necessarily be the most cost-effective approach for certain uses. It is not, despite its reputation as a tool for upwards accountability, widely used to hold organisations to account. Learning is highly dependent on genuine participation – which is not necessarily a feature of many of the time-poor evaluations motivated by administrative ritual and symbolic accountability or characterised by the report and adherence to predetermined methods. The use of many evaluations is confined to particular groups, often excluding the very stakeholders who are key to cross-departmental decision-making, and providing an unlikely basis for organisational learning.

The responsibility of evaluation lies in identifying the strengths and weaknesses of programmes, which it can do extremely well, and in facilitating utilisation, which it has been doing less well. Serious participation and a far greater focus on the intended user and uses would help to expose the practice of inappropriate or ritual evaluation and prevent evaluation further contributing to the current mistrust and saturation in the sector.

Much more effort may also be needed to identify better organisational-change agents to continue the process that accountability tools can start but not complete. If external pressure is a significant change agent, then audits may serve upwards accountability better than evaluations, linked to regulatory bodies designed for the purpose. Participatory and impact evaluation designed with, and for, the affected communities would strengthen downwards accountability. Organisational and individual learning is better served by interactive approaches characterised by openness, not judgement. Isolated efforts to innovate, such as those implemented by ActionAid, can be built upon and shared to make learning and accountability experimental and creative, rather than potentially moribund and mistrusted.

3.5.2 Considerations for the future

The focus of accountability tools should shift from methodology to utilisation. If the goal is to enhance performance, then the door is open to a range of possible tools, of which evaluation is only one. To an extent, different expectations of evaluation derive from the belief system that underpins accountability. Some users and evaluators perceive a wilfully recalcitrant sector that must be forced to change. This view will increasingly lead to an increase in tools and evaluations to measure performance, compliance with standards, accreditation and so on. This is similar to the route already taken by the public sector in many developed countries. Evaluations serving this purpose would be predominantly formal and published. Consideration would have to be given to ensuring that such evaluation teams were given the kind of mandatory access and cooperation that is commonly the preserve of financial audits.

Others view the humanitarian sector as, essentially, populated by highly motivated personnel doing their best in difficult circumstances. This latter viewpoint lends itself to an approach to evaluation that is predicated on trust and the creation of safe spaces for learning, including making mistakes. Either way, the choice of tools that best generate change would be based upon a clear analysis of which approach works best. Some organisations may respond better to the stick and others to the carrot. The impact of evaluation is enhanced, if not enabled, by being part of a broader menu of approaches to enhancing performance. Monitoring, for example, remains a poor cousin of evaluation and has yet to receive the same attention from decision-makers. Evaluation, in whatever form, is only one element of accountability.

If the indications in this study, that evaluation has a relatively poor image with operational staff and senior managers, hold true more broadly, then there is work to be done to reduce the gap between evaluation and users. A weakness of this study was limited consultation with field personnel and senior managers. This is also a weakness of other studies. Given that these groups are, or should be, essential users of evaluation, this is a curious oversight. It would be useful to pursue this line of enquiry, finding out what kind of accountability and learning tools such users would value.

It may be that much more could be done to adapt evaluation to serve organisational learning better. However, evaluation may not be the best tool for this at all; more

appropriate methods may be found in approaches currently used in the private and public sectors. More attention could be paid to what has actually led to enhanced performance, both within specific organisations and across the sector, by examining significant changes in performance and tracing backwards to the causes.

Whatever the accountability tool, no instrument will generate change unless the sector finds the time and space to allow it. At present, there is a sense that the humanitarian sector is suffering from inertia generated by an overload of tools, guidelines and bureaucracy. Evaluation itself may be inadvertently contributing to the workload. The decision to carry out an exercise in learning and accountability should be carefully considered; more robust prioritisation of evaluation by all stakeholders may be necessary now. Part of the menu of tools could simply entail giving staff the time to innovate and absorb information. Perhaps there is just one leaf to take out of Google's book, which is reported to include giving 'each employee a day a week to work on their own projects, the genesis of many of the new services the company is rolling out every other week' (*Observer*, 3 September, 2006).

References

ALNAP (1997) *Organisational and Institutional Learning in the Humanitarian Sector. Opening the Dialogue*. Koenraad van Brabant. London: ODI.

ALNAP (2002) *Humanitarian Action: Improving Performance through Improved Learning*. London: ODI.

ALNAP (2003) *Annual Review 2002*. London: ODI (http://www.alnap.org/AR2002/AR2002_fulltext.pdf).

ALNAP (2004) *Review of Humanitarian Action in 2003. Field Level Learning*. London: ALNAP (http://www.alnap.org/RHA2004/index.html).

ALNAP (2005) *ALNAP Review of Humanitarian Action in 2004. Capacity Building*. London: ODI.

Broughton, B and S Maguire (2005) *Inter-agency Real-time Evaluation of the Humanitarian Response to the Darfur Crisis*. OCHA.

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

Carlsson, C and P Engel (2002) *Enhancing Learning through Evaluation: Approaches, Dilemmas and Some Possible Ways Forward* (background papers). Maastricht: ECDPM (http://www.ecdpm.org/Web_ECDPM/Web/Content/Navigation.nsf/index2?readform&http://www.ecdpm.org/Web_ECDPM/Web/Content/Content.nsf/0/05A45EF48F023482C1256C7E00555A9C).

Carlsson, J, M Eriksson-Baaz, AM Fallenius and E Loevgren (1999) *Are Evaluations Useful? Cases from Swedish Development Co-operation*. Sida Studies in Evaluation (http://www.sida.se/sida/jsp/sida.jsp?d=118&a=2355&language=en_US).

CGD (Center for Global Development) (2006) *When Will We Ever Learn? Improving Lives through Impact Evaluation. Report of the Evaluation Gap Working Group*. May. Copyright © 2006 by the Center for Global Development, ISBN 1-933286-11-3.

Charity Commission (2005) *Report of Findings of a Survey of Public Trust and Confidence in Charities*. Prepared by Opinion Leader Research for The Charity Commission, November.

Church, Cheyanne (2005) *Mind the Gap: Policy Development and Research on Conflict Issues*. INCORE, January (www.incore.ulster.ac.uk).

Corbett, M and A Lloyd (2005) *MSF Holland's Medical and Nutritional Strategy in Darfur – Sudan*. MSF, May.

Edgren, Gus (1999) *Fashions, Myths and Delusion: Obstacles to Organisational Learning in Aid Agencies* In Jerker Carlsson and Lennart Wohlgemuth (eds) *Learning in Development Co-operation*. Stockholm: Almqvist and WIksell International.

Fritz Institute (2006) *Evidence of Impact: Challenges and New Directions*. Impact Conference, 19–20 May, 2006, Sebastopol, California Conference Proceedings.

Gordillo, G and K Anderrson (2004) 'From Policy Lessons to Policy Actions: Motivation to Take Evaluation Seriously' in *Public Administration and Development* 24, pp 305–320 (http://www3.interscience.wiley.com/cgibin/abstract/109629683/ABSTRACT?CRETRY=1&SRETRY=0).

HAPI (2006) *The Accountability Alibi*. N Stockton. Humanitarian Accountability Partnership International, May (www.hapinternational.org).

HPG (2002) *Financing International Humanitarian Action: A Review of Key Trends*. HPG Briefing Paper, ODI, November.

Kent, R (2004) *Humanitarian Futures. Practical Policy Perspectives.* HPN paper number 46, April. London: ODI/ Humanitarian Practice Network.

Landry, Réjean, Moktar Lamari and Nabil Amara (2001) *Extent and Determinants of Utilization of University Research in Public Administration,* 29 July to 1 August, Canada (http:// www.fss.ulaval.ca/kuuc).

Mango (2004) *Helping People is Difficult: A Contribution to Recognising the Constraints and Releasing the Energies of NGO Staff*. A Jacobs, 18 November.

MFA Netherlands (2000) *Evaluation Feedback: Experiences and Recent Changes*. R. van den Berg and T Kliest. Policy and Operations Evaluation Department, Ministry of Foreign Affairs of the Netherlands. Paper prepared for the OECD/DAC Working Party on Aid Evaluation Workshop on Evaluation Feedback for Effective Learning & Accountability, Tokyo, 26–28 September 2000.

MFA Finland (1997) *Utilisation of Evaluation Results in Finlands' International Development Cooperation*. Ministry of Foreign Affairs, Department for International Development Cooperation, Finland, June.

NAO (2005) *Report by the National Audit Office: Engaging with Multilaterals.* National Audit Office, Department for International Development, December.

OCHA (2006) *Lessons Learned in Conducting the Inter-Agency Real-Time Evaluation of the Humanitarian Response to the Darfur Crisis*. Prepared by OCHA's Evaluation and Studies Unit (OCHA/ESU), January.

O'Donoghue, G (2005) *Rudder and Sails. A Review of CAFOD'S Humanitarian Function*. CAFOD, June.

OECD (2000) *Evaluation Feedback: Experiences and Recent Changes*. R Van den Berg and T Kliest, Policy and Operations Evaluation Department, Ministry of Foreign Affairs of the Netherlands. Paper prepared for The OECD/DAC Working Party on Aid Evaluation.

OECD (2001) *Evaluation Feedback for Effective Learning and Accountability* (http://www.oecd.org/dataoecd/10/29/ 2667326.pdf).

O'Reilly, F and J Shoham (2002) *MSF Holland in Afghanistan Mission Evaluation: May 2000 – May 2002*. MSF.

Patton, M (1997) *Utilization-Focused Evaluation*. London: Sage.

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 (http://www.alnap.org/pubs/pdfs/ followup.pdf).

Ramalingam, B (2005) *Implementing Knowledge Strategies: Lessons from International Development Agencies*. Working Paper 244, ODI, April.

Ramalingam, B (2006) *How Can Outcome Mapping Help Strengthen Knowledge and Learning Strategies?* Presentation given at Crossing Borders, Crossing Boundaries, 2005 Joint AEA/CES Conference, 24 to 29 October 2005 (downloaded from http://www.idrc.ca/ en/ev-91687-201-1-DO_TOPIC.html).

Ridde, Valéry and Shakir Sahibullah (2005) 'Evaluation Capacity Building and Humanitarian Organization' in *Journal of Multi Disciplinary Evaluation (JMDE)*, 3 October.

Sida, L and Chris Szpak (2004) *An Evaluation of Humanitarian Information Centers including Case Studies of HICs*

for Iraq, Afghanistan, and Liberia. USAID/ DFID.

Shulha, L, M Cousins and J Bradley (1997) 'Evaluation Use: Theory, Research and Practice since 1986' in *Evaluation Practice* 08861633, Vol 18, Issue 3, Fall.

Slim, H (2002) 'By What Authority? The Legitimacy and Accountability of Non-governmental Organisations' in *The Journal of Humanitarian Assistance,* 10 March (http://www.jha.ac/articles/ a082.htm).

Suhrke, A (1999) 'From One Crisis to Another: Organisational Learning in UNHCR' in Jerker Carlsson and Lennart Wohlgemuth (eds) *Learning in Development Co-operation*. Stockholm: Almqvist and WIksell International.

Telford, J and J Cosgrave (2006) *Joint Evaluation of the International Response to the Indian Ocean Tsunami: Synthesis Report*. London: Tsunami Evaluation Coalition.

UNHCR (2001) *Real-time evaluation of UNHCR's response to the Afghanistan emergency*. Bulletin No. 2, J Crisp and E Stigter.

Valid International (1999) *MSF in Catastrophe*. Steve Collins, Valid International.

Weiss, Carol (1977) 'Research for Policy's Sake: The Enlightenment Function of Social Research' in *Policy Analysis* Vol 3, No 4, pp 531–545.

Wigley, Barb (2005) *The State of UNHCR's Organization Culture*. EPAU, UNHCR.

Wilding, J, Issaka Idrissa Mossi, Debra Lynne Edwards, Amanda Weisbaum, Boube Aw, Timothy Mander (2005) *Joint Independent Evaluation of the Humanitarian Response of CARE, CRS, Save the Children and World Vision to the 2005 Food Crisis in the Republic of Niger.*

Williams, K, B de Laat and E Stern (2002) *The Use of Evaluation in the Commission Services – Final Report* (http:// europa.eu.int/comm/budget/evaluation/ pdf/use_of%20_evaluation_final_ report.pdf).

WFP (World Food Programme) (2005) *Study on WFP Follow-up Evaluation Recommendations*. Executive Board Summary (Document Number: WFP/ EB.A/2005/7-B) (http:// documents.wfp.org/stellent/groups/ public/documents/eb/wfp050928.pdf).

Wright, Malaika and Pauline Wilson (2006) 'Can Joint Evaluations Promote Ongoing Collaborative Action by NGOs? (Emergency Capacity Building Project)' in *Humanitarian Exchange* No 34, June, ODI/ Humanitarian Practice Network.