Chapter 3

Joint evaluations coming of age? The quality and future scope of joint evaluations

Tony Beck and Margie Buchanan-Smith

3.1 Introduction

3.1.1 Background and purpose of the meta-evaluation

This is the sixth ALNAP meta-evaluation of evaluations of humanitarian action (EHAs), continuing the series started in 2001. This longitudinal assessment of the quality of EHA is perhaps the most detailed meta-evaluation in development assistance to date. The overall aim of the meta-evaluation is to improve evaluation practice by identifying areas of weakness, that deserve attention, and examples of good practice that can be built upon. There is qualitative evidence that this aim is being met. Lipsey (2000) defines a meta-evaluation as 'meta-analysis and other forms of systematic synthesis of evaluations providing the information resources for a continuous improvement of evaluation practice'. The ALNAP approach is to review periodically a sample of evaluation reports against a Quality Pro Forma that has been developed according to accepted evaluation good practice. Systematic use of the Pro Forma over a number of years has made it possible to identify trends in evaluation quality over time.

3.1.2 The focus of this year's meta-evaluation: joint evaluations

For the last five meta-evaluations, the sample has been dominated by single-agency evaluations, the most usual form in which EHAs are carried out. Tracking changes in quality and approach over time has thrown up some interesting results. For example, the use of DAC criteria in EHA has gradually strengthened in the last few years, and consultation with primary stakeholders has improved in evaluation methodology. There has been little or no improvement in other areas, however, such as attention to the crosscutting issues of gender equality, protection and advocacy. Where improvement in quality has been evident, the improvement has usually happened quite gradually. This begged the question of whether another meta-evaluation in 2007, again dominated by single-agency evaluations, would yield new insights and contribute to improved evaluation practice.

Meanwhile, there has been a growing trend towards 'jointness' in the aid world, and joint evaluations (JEs) of humanitarian action.[1] Originally championed by donor

governments, joint evaluations have been the focus of recent and growing interest and engagement from NGOs and UN agencies. In 2005/06 the second-ever system-wide evaluation of humanitarian action took place, of the international response to the Indian Ocean tsunami of 2004, through the Tsunami Evaluation Coalition (TEC).[2] Some of the reasons for this growing interest in joint evaluations are explored below. The sector is on a steep learning curve in terms of how to do joint evaluations, including how best to manage and organise them, when they are appropriate, and with whom. This is accompanied by an active debate about the pros and cons of joint evaluations, and how they relate to single-agency evaluations – for example, can they replace them?

The current meta-evaluation provides a timely opportunity to focus on recent joint evaluations of humanitarian action, and to contribute to debate and practice about how to strengthen future joint evaluations. The specific objectives of this year's meta-evaluation are:

1 to review the quality of joint-evaluation exercises, where possible comparing this with the quality of past single-agency evaluations

2 to document in an accessible way some of the learning from the growing experience of joint evaluations – especially examples of good practice – to feed into future joint endeavours

3 and thus to make a significant contribution to the nascent but currently limited body of knowledge about joint evaluations.

It is not intended to provide a guide on 'how to do joint evaluations'. There are a number of other publications designed to achieve that purpose (see Section 3.2.3 below). For the purpose of this meta-evaluation, we have defined joint evaluations as evaluations carried out by two or more agencies, evaluating the work of two or more agencies.

3.1.3 Methodology and sample

The methodology used is similar to that of previous meta-evaluations, with some minor adjustments to take account of the joint-evaluation focus. A sample of 18 joint

evaluation reports was selected for this meta-evaluation, and is detailed in Annexe 3.5.[3] The meta-evaluators[4] examined joint-evaluation quality through assessment against the Pro Forma (slightly amended to ensure its relevance to joint evaluations), and evaluation process through interviews with those involved in joint evaluations, ensuring iteration between these two methods. The data from the assessment against the Pro Forma were analysed and compared with results from previous ALNAP meta-evaluations, which have covered a total of 138 evaluations.

Interviews were held with 22 representatives from 15 different organisations: 5 UN agencies, 5 NGOs (or NGO bodies), ECHO and the ALNAP Secretariat. The interviews focused on the purpose, planning and management of the joint evaluation, and on the post-evaluation process and utilisation. Six evaluators who had led or been centrally involved in one or more of the joint evaluations in the sample were also interviewed by telephone or in person. These interviews similarly focused on the purpose and management of the evaluation and on the post-evaluation process, but also explored evaluation methodology. The full methodology for this meta-evaluation is described in Annexe 3.1, and the Pro Forma is presented in Annexe 3.2. Annexe 3.3 lists all those interviewed for this meta-evaluation, and Annexe 3.4 is the questionnaire used for agency interviews.

In order to guide our work in this meta-evaluation, a number of working hypotheses were drawn up at the outset, to be tested and explored. These are presented in Box 3.1; our findings against these hypotheses are presented throughout the report, in the sections indicated in Box 3.1.

Box 3.1 Working hypotheses to guide the 2006/07 meta-evaluation

1. The experience of joint evaluations helps to build trust and social capital within the sector (Section 3.3.2).
2. Joint evaluations tend to be driven from the centre (ie headquarters) rather than from the field (Section 3.3.3).
3. Involvement of the government of the area affected by the humanitarian crisis is still weak in humanitarian joint evaluations (Section 3.3.3).
4. There is greater opportunity for beneficiaries to be consulted/surveyed in joint evaluations than in single-agency evaluations (Section 3.4.1).

CONTINUED

Box 3.1 Working hypotheses to guide the 2006/07 meta-evaluation *continued*

5 Joint evaluations have more rigorous methodologies than do single-agency evaluations (Section 3.4.2).

6 Joint evaluations pay more attention to international standards and guidelines than do single-agency evaluations (Section 3.4.3).

7 Joint evaluations are stronger than single-agency evaluations on crosscutting issues such as gender and protection (Sections 3.4.5 and 3.5.5).

8 The overall quality of joint evaluations tends to be higher than that of single-agency evaluations (Section 3.4.6).

9 Joint evaluations are more likely than single-agency evaluations to address both policy issues and programme performance (Section 3.5).

10 Joint evaluations pay attention to wider debates within the humanitarian sector, and situate their findings accordingly (Section 3.5).

It could be argued that a sample of 18 joint evaluations is rather small as a basis from which to draw conclusions about the 'state of the art'. However, only 25 joint-evaluation reports were provided to the ALNAP Evaluations Report Database (ERD) for 2005–2007 (see below), and our sample is representative of the system and in particular its main joint-evaluation initiatives over the last three years; it also allowed us to assess the key issue of evaluation process.[5] Rather than providing the last word on joint evaluations, this meta-evaluation is intended to take stock, at a key moment when interest and investment in joint evaluations appears to be on the rise, and there is a hunger for learning and guidance on what is working, how they compare in quality with other types of evaluation, and especially how they can be improved. This context emerged strongly in many of the agency interviews. And of course a sample of 18 joint evaluations can provide much greater coverage of both agencies and programmes than a sample of 30 single-agency evaluations, precisely because a number of agencies have come together in each evaluation exercise. Perhaps more challenging has been the attempt to include both real-time evaluations (RTEs) and ex-post evaluations in one meta-evaluation exercise. As noted below, these can differ quite substantially in approach and output, which raises a question about whether the current Quality Pro Forma, as used for the meta-evaluation, is appropriate for RTEs.

The decision to omit certain evaluation reports in selecting the sample was guided by the following three considerations.

1 As in previous meta-evaluations, evaluation reports that deal overwhelmingly with institutional issues were omitted, as the Pro Forma has been designed to assess evaluation reports dealing with humanitarian responses. Thus, the UN Pro-Cap evaluation was omitted.[6]

2 Although a small part of the sample consists of joint RTEs, these are rather different from ex-post joint evaluations and may therefore be best assessed against an adapted Pro Forma in a future meta-evaluation. Four reports (just over 20 per cent of the sample) are RTEs, and at least two joint RTEs on ALNAP's database were omitted.

3 While the second-ever system-wide joint evaluation – the TEC – deserved prominence in this meta-evaluation, it was agreed that only four of the six TEC reports should be included, to avoid undue skewing of the sample towards the TEC. The four TEC reports included therefore also represent just over 20 per cent of the sample.

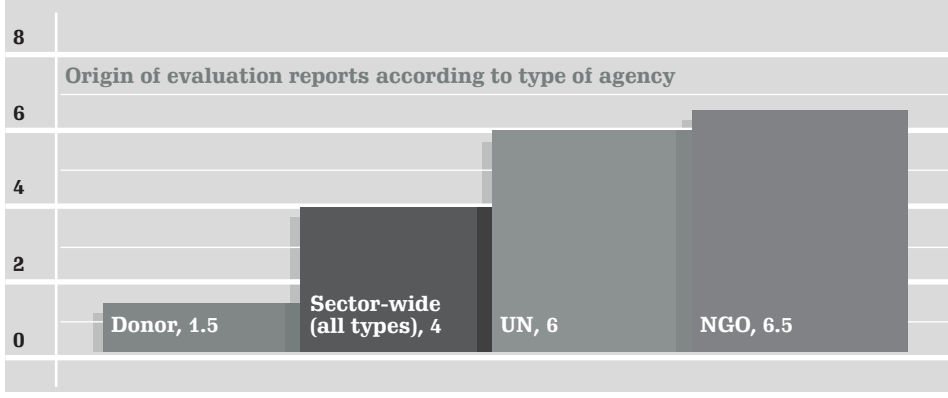
The breakdown of the sample according to different types of agency is presented in Figure 3.1. Joint evaluations initiated by UN agencies and NGOs clearly dominate. As explained in more detail below (Section 3.2.2), there has been growing interest in joint evaluations in both of these camps. In this meta-evaluation we have paid particular attention to a new NGO initiative around joint evaluations, as part of the Emergency Capacity Building Project (ECB), also explained below, and the sample includes all five of this project's joint evaluations. In short, we have evaluations from four main sources: the Interagency Health Evaluation initiative (IHE), the TEC, ECB and the Inter Agency Standing Committee (IASC). All of these are rich experiences that offer valuable learning for the future of joint evaluations.

It is striking that so few donor agencies have been the initiators of joint evaluations in the last couple of years, despite being the early champions of joint evaluations. Several donor agencies have, however, played a prominent role in the TEC. In previous meta-evaluations the European Community Humanitarian Office (ECHO) has been heavily represented (Wiles, 2005). However, ECHO is a relative newcomer to joint evaluation, and contributes just one evaluation to this set. The absence of the

Red Cross Movement from the sample is also striking; it has not lodged any joint evaluations on ALNAP's ERD for the period under review. (As noted in previous meta-evaluations, the decision to submit evaluation reports to ALNAP's ERD is an entirely voluntary one. Thus, the composition of the ERD may be indicative of EHA output in the sector, but it is not exhaustive. For example, contributions from national governments and national NGOs rarely appear.)

Over 80 per cent of reports in the sample are joint evaluations of responses to natural disasters, and especially sudden-onset disasters. This is not surprising considering that 2005–2007, the period in question, was dominated by the massive international response to the Indian Ocean tsunami, in turn generating a lot of evaluation activity. Almost half of the natural-disaster evaluation reports in the sample are to do with the tsunami response. Joint evaluations can be demanding exercises to get off the ground and to implement, especially if there are many different agendas to be considered and harmonised. This begs the question about whether it is easier to reach agreement to undertake a joint evaluation after a natural disaster, rather than in relation to a conflict in which political agendas may be stronger and more sensitive. (However, the distinction between natural disaster and complex emergency is not always easy to make, as illustrated by conflict in tsunami-affected parts of Aceh and Sri Lanka.)

Figure 3.1 Breakdown of meta-evaluation sample according to origin of report (number of reports)



3.1.4 A guide to this chapter

Section 3.2 begins with a brief review of the history of joint evaluations, exploring the motivations for the growing number of joint evaluations of humanitarian action in the last few years. On this basis a typology for joint evaluations is proposed, closely related to the purpose of joint evaluations. Section 3.3 considers the different aspects of setting up and managing a joint evaluation, ranging from negotiating and drawing up the terms of reference (ToR), to management roles and responsibilities and selecting the evaluation team. Section 3.4 assesses the quality of joint evaluations, against the Quality Pro Forma. Sections 3.3 and 3.4 present most of the data analysis for this year's meta-evaluation, where possible making comparisons with the results of previous meta-evaluations.

Section 3.5 investigates the extent to which joint evaluations situate their findings in wider policy debates, helping to move those debates forward. Section 3.6 explores the post-evaluation process, looking at the accessibility of evaluation reports, their conclusions and recommendations, and follow-up mechanisms. Section 3.7 provides a concluding analysis of the quality of joint evaluations, summarises some of the main learning about doing them, and proposes a future agenda for joint evaluations of humanitarian action. Because of the importance of the TEC, we have included a separate item (Box 3.2) on the strengths and weaknesses of this joint evaluation.

3.2 The evolution and purpose of joint evaluations

This section looks at the origins and early development of joint evaluations, and includes a note on the existing literature on joint evaluations. Different types of joint evaluations are discussed, and an expanded typology is presented, which may be useful in the process of planning joint evaluations. Finally, the purpose – or purposes – of joint evaluations are considered. These are typically 'learning' and/or 'accountability'. Findings from analysis of the meta-evaluation sample are presented here, illustrating the range of different purposes found within the sample.

3.2.1 The origins of joint evaluations in the development sector

Joint evaluations have a longer history in development than in the humanitarian sector. They were originally pioneered by donor governments coming together through the Development Assistance Committee (DAC) of the Organisation for Economic Co-operation and Development (OECD), with the somewhat narrow administrative focus of improving understanding of different donors' procedures and agendas (DAC, 2005a). However, this aim has broadened over the years as the dominant mode of aid assistance has shifted to budget support and sector-wide approaches, usually through multi-donor programming, clearly strengthening the rationale for doing joint evaluations. The importance of involving aid recipients in a joint evaluation has long been recognised, but their role appears to be shifting in the development sector from fairly passive to more pro-active, as the emphasis switches from 'multi-donor' to 'multi-partner'. The Paris Declaration on Aid Effectiveness reinforces this shift, although there is still a long way to go before recipient governments are truly partners (DAC, 2005b).

3.2.2 Evolution of joint evaluations in the humanitarian sector

Widespread interest in joint evaluations in the humanitarian sector is much more recent. The first significant experiment in a joint evaluation in the humanitarian sector was the seminal multi-agency Rwanda evaluation in 1996 (Borton et al, 1996). This experience was well-documented[7] and has continued to inform debate and practice on joint evaluations as well as boosting interest in programme evaluations.

In the following years, most examples of joint evaluations arose through donor governments coming together to evaluate: humanitarian action in a particular country (for example five donors evaluating humanitarian assistance and reconstruction in Afghanistan in 2005); the performance of a particular agency or group of agencies (for example the Dutch and British governments evaluating WFP's programme in Sudan in 1999); or a thematic issue (for example a number of donors coming together to evaluate humanitarian assistance to internally displaced persons in 2005). The Nordic governments have been the main pioneers in many of these cases.[8]

During the last decade there has been a growing trend of UN agencies coming together to carry out joint evaluations of their combined work in a particular country or region. To some extent donors have encouraged this, but the motivation has also come from within the UN system. In 2001, for example, WFP and UNHCR came together to evaluate 'protracted relief and recovery operations' in Sudan, in which the two agencies had been cooperating closely under an MoU. In 2005/06, the two agencies came together again to evaluate pilot food-distribution projects in five countries in which WFP had taken over responsibility from UNHCR for food distribution to refugees and internally displaced persons (IDPs). (A report of this evaluation is included in the meta-evaluation sample.)

At a different level, and with different motivations, WHO and UNHCR took the initiative in 2003 to launch inter-agency health evaluations (IHEs), bringing in a number of other actors including NGOs and academic institutions. The motivation here was to capture what was happening sectorally, in terms of the bigger picture, as it was accepted that this could not be addressed in single-agency evaluations but is an important contribution to learning, especially at the policy level.

Joint RTEs are an even more recent phenomenon in the UN system. One of the first was the 'Inter-Agency Real Time Evaluation of the Humanitarian Response to the Darfur Crisis', commissioned by the UN Emergency Relief Coordinator and Under-Secretary-General for Humanitarian Affairs in 2004, and published in 2006. Aware of the criticism that the humanitarian response had been 'woefully inadequate', the motivation was to benefit from external guidance in order to improve the operational response in real time, and to identify broader lessons applicable elsewhere.

> The approach was groundbreaking in that it was the first attempt to comprehensively evaluate an ongoing crisis across all sectors and functions using a participatory approach involving all key stakeholders *while* the response was still underway. (Broughton and Maguire, 2006, p 1)

The IASC has continued to pilot inter-agency RTEs, three of which are in our sample. A more recent motivation is to monitor the progress and impact of the UN-led humanitarian reform programme on the ground: what difference is it making? The UN Office for the Coordination of Humanitarian Affairs (OCHA) has been charged with managing these RTEs. As noted by an OCHA staff member, this throws up the challenge of pioneering two new approaches to evaluation – real-time and joint – at

the same time. Meanwhile the United Nations Evaluation Group (UNEG) is giving renewed emphasis to joint evaluations on the development side, as part of the 'delivering as one UN' agenda (UNEG, 2005).

Until recently the Disasters Emergency Committee (DEC) in the UK had been the main pioneer of joint evaluations in the NGO sector, evaluating the use of appeal funds raised collectively by the (currently 13) member NGOs. These joint evaluations have been ongoing for about a decade, and have been cited as examples of good practice in previous meta-evaluations, for example the 2004 meta-evaluation, for their clear and comprehensive terms of reference, and their ability to weave together principles such as The Red Cross/Red Crescent Code of Conduct with the DAC criteria (Wiles, 2005). However, it is unlikely that these joint evaluations will continue as the DEC develops a new accountability framework which will probably no longer include joint evaluations, favouring instead collective monitoring missions, for learning rather than accountability purposes.

Meanwhile, the joint-evaluation baton has been picked up by a group of seven international NGOs – members of the Interagency Working Group – that created the Emergency Capacity Building (ECB) project beginning in 2005.[9] Early in the project they pioneered a number of joint evaluations, starting with evaluations of their response to the Indian Ocean tsunami. Joint evaluations have been a prominent element of this first phase of the ECB project, and a total of five have now been completed, in Asia, Africa and Central America (all included in our sample). There appear to have been a number of motivations for launching joint evaluations within the ECB project. To some extent it was seen as a way of avoiding a plethora of different evaluation processes, especially when the international response to the tsunami had been so crowded; it was also a way of making good use of scarce evaluation resources. A second motivation was to strengthen accountability, with a strong emphasis on downward accountability and the importance of capturing the views and perspectives of local affected people. The capacity-building element of the ECB project highlighted learning, both peer learning programmatically across agencies and also learning about doing joint evaluations and in the process building the evaluation capacity of ECB member agencies.

The European Commission is no newcomer to joint evaluations: a joint EU evaluation of programme food aid was commissioned as long ago as 1993. But joint evaluations are a new phenomenon for ECHO, which until recently favoured single-

agency and single-programme evaluations. This is beginning to change. ECHO's evaluation department sees joint evaluations as the trend for the future, with partner agencies it is funding (eg UN agencies) and increasingly with member states. The new 'European Consensus on Humanitarian Aid', recently signed by the Presidents of the European Commission, Council and European Parliament in December 2007, specifically promotes 'joint approaches' to evaluations by donor governments (EU, 2007, p 10). In 2006 ECHO launched a series of joint real-time evaluations with WHO and some donors (and one report of one of these is included in this meta-evaluation sample). Similar to the IASC experience, this is pioneering both real-time and joint approaches in one exercise. The joint nature of the evaluations is providing valuable learning opportunities across participating agencies, while the real-time element is becoming popular with operational departments because of the immediate feedback it provides.

After the demanding but much-feted sector-wide Rwanda evaluation, it is interesting to note that it took a further ten years before another system-wide evaluation was launched – by the Tsunami Evaluation Coalition. On at least a couple of occasions the idea of launching a system-wide evaluation had been discussed, usually among the ALNAP membership, for example to evaluate the response to Hurricane Mitch in late 1998. There was an expectation that as a natural disaster this would be easier and less sensitive than the Rwanda crisis, but no donor agency was prepared to take the lead role that Danida had played in the Rwanda evaluation. The prospect of a system-wide evaluation was raised again within ALNAP in relation to the Kosovo crisis, but on this occasion the political and military sensitivities meant that donor governments were reluctant to take it on. The enormity of the tasks of launching and leading a system-wide evaluation also appear to have been intimidating factors.[10] Eventually it took the tsunami disaster, and the unprecedented scale of the response and funding, to trigger the second system-wide evaluation. In the words of one interviewee: 'there was a sense that it was high time to launch another system-wide joint evaluation'.

Thus, in the mid-2000s, there is a growing momentum behind the joint-evaluation approach. To some extent this is to do with a change in how the sector is operating, with greater emphasis on working together. Strikingly, most parts of the humanitarian sector are now engaged: donors, UN agencies and NGOs. The most obvious gap is recipient governments, national NGOs and research institutions (discussed below). But there is a sense that the commitment to joint evaluations is still quite fragile, apparent at the workshop on joint evaluations at the 20th ALNAP

biannual meeting in Rome in December 2006. There is clearly a commitment and interest in moving the agenda forward, as frequently expressed by interviewees for this meta-evaluation. However, attempts to institutionalise joint evaluations have been mixed. On the one hand, the DEC and IHE are letting them go; on the other hand the NGOs engaged in the ECB project still seem committed, pending funding for the second phase of this project, as does the IASC for its joint RTEs. Generating and maintaining interest in joint evaluations institutionally has usually required persuasive and influential champions.

3.2.3 The literature on joint evaluations

The literature on joint evaluations is still sparse, reflecting the relative infancy of this approach. Not surprisingly, the DAC has published the most comprehensive documentation on joint evaluations, although most of this is to do with government participation, whether donor or recipient governments; there is little mention of NGOs or UN agencies. A few articles and papers for conferences have been written on individual joint-evaluation experiences,[11] and there are also a number of 'how to' guides on joint evaluations, produced by donor, UN and NGO agencies respectively.[12] Much of this literature is preoccupied with experiences and learning about the management of joint evaluations, reflecting the complexities and difficulties of management as a joint process (as discussed in Section 3.2 below). In comparison, there is little on evaluation methods.

Almost every document written on joint evaluations rehearses the arguments for and against, often at some length.[13] There is a preoccupation with the questions, 'Are they worth it?' and 'When are they appropriate?'. The useful DAC publication taking stock of joint evaluations (DAC, 2005a) groups the reasons for doing joint evaluations into five categories.

1. Overarching policy reasons: the benefit that a joint evaluation provides in 'seeing the big picture', and evaluating the programme or range of interventions against this big picture is frequently mentioned in much of the literature.

2. Evaluation strategy motives: for example, to do with the increased credibility and legitimacy that joint evaluations can provide, which can be useful in advocating for change, especially if there are sensitive issues to be covered.

3 Learning motives: so partners understand each other's approaches and exchange good practice.

4 Managerial, administrative and financial motives: for example, sharing funds if evaluation resources are scarce, or redressing a lack of sufficient evaluation capacity within an agency.

5 Developmental motives: for example, reducing transaction costs for developing countries, and building ownership and participation of developing countries (although this latter point has been a stronger motivation in the development than the humanitarian sector so far).

The reasons against doing joint evaluations are often to do with the complexity of the process and of the subject, which can result in a time-consuming and expensive project, not least in transaction costs,[14] involving complicated management structures.

3.2.4 Categorising joint evaluations

There are many different ways to classify joint evaluations. The DAC review (2005a, p 16) proposes a simple typology in the hope that this will contribute to greater analytical rigour, and avoid confusion and misunderstanding when partners work together. The DAC typology is:

1 classic multi-partner – participation open to all stakeholders

2 qualified multi-partner – participation open to those who qualify as part of a particular group

3 hybrid multi-partner – reflecting more complex ways of working, eg with some partners taking a less-active role.

Some of the most useful ways of categorising joint evaluations seem to be around two criteria: purpose and scope of the evaluation, and how actors work together[15] (as captured in the DAC typology). Building on the DAC typology, we developed a categorisation to fit the humanitarian sector, and have plotted the 2006/07 meta-evaluation sample against this typology (Table 3.1).

Table 3.1 Proposed typology for joint evaluations in the humanitarian sector (with examples from the meta-evaluation sample)

HOW ACTORS WORK TOGETHER		FOCUS OR SCOPE OF EVALUATION		
PROGRAM FOCUS	INSTITUTIONAL	SECTORAL OR THEMATIC FOCUS	MULTISECTORAL FOCUS, RELATED TO A PARTICULAR HUMANITARIAN CRISIS (USUALLY BOUND GEOGRAPHICALLY)	GLOBAL (eg, GLOBAL POLICY)
'PARTNERSHIP' Donor & recipient agencies evaluate together as equal partners				
		ECHO/WHO/DFID JE:WHO emergency response, Pakistan		
'LIKE-MINDED AGENCIES' (OR QUALIFIED) Agencies with similar characteristics coming together				
		WFP/UNHCR pilot food distribution (UN agencies operating to a MoU)	**All ECB evaluations (groups of NGOs); DEC evaluations (groups of NGOs); IASC RTEs (UN agencies)**	
'HYBRID MULTIPARTNER' Disparate actors coming together, playing variable roles (eg, active/passive)				
		IHE evaluations (comprising UN agencies, NGOs, academics, recipient government, etc)		
'SYSTEM-WIDE' Open to all actors in the system				
			TEC evaluation	

What emerges is that most joint evaluations of humanitarian action (at least in the 2006/07 sample) are multi-sectoral, focused on a particular humanitarian crisis. However, there are different ways in which agencies have come together to carry out these evaluations, the most popular configuration is that of 'like-minded agencies', whereby agencies with similar characteristics/background come together, in this case a grouping of international NGOs, or a grouping of UN agencies. All other evaluations in the sample are sectoral or thematic. It is also interesting to note the blank columns in the Table 3.1. There may have been institutional joint evaluations carried out in the last couple of years that have not been shared with the ALNAP ERD.[16] But to the authors' knowledge there have been no joint evaluations looking at global policy issues (the final column of Table 3.1). It may be that this is seen as the domain of research rather than jointly undertaken evaluations.

With reference to the rows in the table – how actors work together – it seems reasonable to expect that the joint evaluation will be increasingly complex to manage as one moves down the spectrum from 'partnership' to 'system-wide'. Thus, one would expect that a system-wide, multi-sectoral joint evaluation like the TEC would be considerably more challenging than a 'like-minded agencies', multi-sectoral joint evaluation such as one of the ECB evaluations. The most challenging joint evaluations are likely to be those in the bottom right-hand corner of the table, with a broad multi-sectoral or global focus and open to all actors in the system. When agencies are considering embarking on a joint evaluation, this categorisation, which relates to complexity, may be a useful reference in deciding what type of joint evaluation to choose.

3.2.5 The purpose of joint evaluations

Accountability and/or lesson-learning are the most commonly stated purposes of evaluations. The tension in trying to achieve both in one exercise has long been recognised: 'For the most part, it is regarded as a creative tension that contributes to the uniqueness and potential value of the evaluation process' (ALNAP, 2001, p 23). In practice, what frequently happens is that one purpose dominates, even if the evaluation is supposed to give equal weight to both. The results from this meta-evaluation mostly confirm this pattern.

The terms of reference (ToR) for most of the joint evaluations in the sample emphasise both accountability and learning. Most of the ECB evaluations are particularly clear on this. The only evaluations that give a strong steer towards learning (and not accountability) are the four real-time evaluations and the IHE evaluations. The TEC evaluations, however, are surprisingly inconsistent in their stated purpose. The coordination study emphasises learning with no mention of accountability; the local and national capacities study mentions both learning and accountability; the ToR for linking relief, rehabilitation and development stress learning as the key purpose although acknowledging the accountability purpose of the whole exercise; but the ToR for the synthesis report emphasise learning with no explicit mention of accountability.

In practice, learning seems to have won through as the dominant purpose of several of the joint evaluations reviewed, and this was confirmed in many of the agency interviews. This was especially evident for the TEC. Dropping the proposed thematic study on impact of the international response to the tsunami disaster was a major blow to achieving real accountability.[17] In addition, as one respondent noted, when the TEC budget had to be cut, it was translation and field-level workshops that were sacrificed. Interestingly, when a survey was carried out among TEC stakeholders after the evaluations had been completed, the results showed very clearly that learning was their first-ranked motive for being involved in the TEC (as cited by 92 per cent of survey respondents); in contrast, accountability ranked fifth (cited by 48 per cent of survey respondents) (TEC, 2007). The usefulness of joint evaluations for learning is a theme in much of the literature, ranging from learning about partners' approaches, to sharing good practice, to learning about the programme or initiative being evaluated.[18] Both the TEC and ECB clearly identified another learning purpose: to learn from the process of implementing a joint evaluation.

It is much more likely that reports from joint evaluations will end up in the public domain, compared with single-agency evaluations, thus fulfilling at least one accountability criterion. This was the case for all the ECB evaluations, even though some ECB member agencies did not follow this practice for their own individual-agency evaluations. Indeed, all evaluation reports in the sample, with the exception of the DEC reports, are publicly available, usually on the Internet. Although the DEC used to put its full evaluations in the public domain, it now releases a much shorter summary version and the full version is treated as a confidential internal document, which may have implications for its accountability to the public and donors.

Another way in which joint evaluations achieve accountability is through peer accountability, especially for joint evaluations between 'like-minded agencies'. The benefits of this were stressed by NGOs participating in the ECB. This is also a feature of joint evaluations by different UN agencies (as described in Section 3.4). Working together requires a level of transparency that is not guaranteed in single-agency evaluations. However, attribution of failure or incompetence to a named agency was not a strong feature of any of the reports in the meta-evaluation set; usually, only success or good practice were specifically attributed. To some extent this limits the accountability function, and is an interesting difference from Study 3 of the joint evaluation of emergency assistance to Rwanda, which was remarkably direct in pointing out incompetence and naming agencies (Borton et al, 1996). But even in that case, there was an interesting discussion about the extent to which it was accountability-oriented in terms of attributing blame to a particular organisation for failure to prevent the very high mortality rates among the refugee population in Goma (Borton, 2001). In the end, the Rwanda study stopped short of this, instead commenting on poor performance, thus indicating the practical limits that joint evaluations are likely to face in meeting accountability objectives.

Some other purposes of joint evaluations have emerged during the process rather than being identified at the outset. This was the case with the ECB. An evolving purpose has been to build evaluation capacity, especially of ECB members that had weak evaluation cultures and were able to learn from more experienced peers. Staff members within these agencies also used the joint-evaluation experience to lobby internally for the evaluation function to be taken more seriously. One reason why this was possible in the ECB was because of the institutionalised nature of the relationship, which meant that a number of joint evaluations were carried out over a two-year period; thus, a longer-term institutional purpose could be realised. A longer-term institutional framework is also a feature of the IHE and IASC joint-evaluation experiments.

Some ECB-agency staff members have emphasised the relationship-building aspect of doing joint evaluations together. This has encouraged a joint approach in other areas as well, for example in risk-reduction work and in carrying out needs assessments.

3.3 Findings on joint evaluation set-up and management

This section presents findings related to joint-evaluation process and management. It is based on assessment of the sample joint evaluations against the ALNAP Quality Pro Forma, agency and evaluator interviews, and document review.

3.3.1 Quality and negotiation of the terms of reference

Developing clear, usable terms of reference (ToR) is a key step in any evaluation. Doing this for joint evaluations offers particular challenges, because the ToR have to be negotiated among multiple parties, some of whom may have conflicting interests. Findings on the quality of joint-evaluation ToR, and comparison to the 2004 meta-evaluation, can be found in Table 3.2.[19] The results show considerably improved practice in 2006/07, perhaps a result of the multi-agency review process. Guidance on developing ToR for joint evaluations is available in DAC (2006) and ECB (2007), and several evaluations referred to the ALNAP Pro Forma, which has supported improved ToR quality.

Table 3.2 Pro Forma area 1.1, The terms of reference (%)

	2004	2006/07
Good	12	38
Satisfactory	34	50
Unsatisfactory	25	6
Poor	29	6

Other key points on ToR are as follows.

- There was strong emphasis on community consultation, for example in the ECB evaluations, the WHO/UNHCR Liberia evaluation, and the WFP/UNHCR food-aid pilot evaluation. This community involvement in developing ToR seems likely to have supported improved community consultation during the evaluation (see Section 3.7 below).

- An emphasis on confidentiality and dignity of respondents was evident in the ECB evaluations, which is unusual and can be considered good practice.

- As in previous meta-evaluations,[20] we found inadequate attention to identification of use and users. Only 6 of the 18 evaluations in the 2006/07 sample fully identified use and users.

The TEC 2005 After Action Review (AAR) found establishing ToR for this large joint evaluation challenging:

Time was not found, or taken, to coordinate the ToR of the different studies. The result of this has been a lack of collective ownership of the ToR. There should have been established early on a tighter strategic policy with each theme taking shape around agreed parameters. ToR should have been tighter, and more tightly bound together by referencing each other.[21]

3.3.2 Management of joint evaluations

Our literature review suggested several management areas for exploration, and we identified four key measures against which to assess our sample.

Time and scheduling

The joint-evaluation literature (eg DAC, 2005a; World Bank, 2005; ECB, 2007) suggests that more time is needed for joint evaluations than single-agency evaluations, because of their broader scope and greater transaction costs. IHE (2007) and ECB (2007) refer to the need for a 30-day evaluation, with the former including a pre-visit for evaluators. Our impression was that more time is not necessarily being allocated for joint evaluations than single-agency evaluations, and that agencies are underestimating the time needed for joint evaluations, both for the evaluation process and for setting up the joint evaluation. For example, a period of 30 days was allocated for the TEC synthesis report, despite its scope, an underestimation that proved problematic. Of the 14 evaluations in our sample that noted constraints, 5 included time as a major constraint. Having said this, there may be a tension between the longer timeframe required for joint evaluations and the difficulties of engaging personnel from agencies, national government and local institutions over an extended duration.

Management bodies and leadership

The DAC (2005) workshop on joint evaluations in development concluded that multi-agency joint evaluations should include a larger steering group, a smaller management group, and involve participation from host countries.[22] Dabelstein (1996) distinguishes between a one-tier management system (with a management committee and contracting delegated to one agency) and a two-tier management system (with steering committee and management committee, for example the joint Rwanda evaluation).[23] A variety of management structures were established for the joint evaluations in our sample (as detailed in Annexe 3.8). Almost all of these evaluations were guided by either an HQ or country-based steering committee, or both.

As well as a core management group in Geneva, the IHE evaluations created a steering group in country, to create ownership, with responsibility for drafting the ToR, managing the evaluation on the ground and devising/implementing a follow-up action plan. Host-country involvement in the steering group was planned, but whether this happened depended upon the national level of involvement in existing health coordination mechanisms. The second IASC Pakistan RTE established an in-country steering committee in order to attempt to overcome some of the resistance to the first Pakistan RTE, a strategy that was partly successful. The TEC experience demonstrated that getting the management structure right is key to supporting a positive evaluation process and results; lack of an adequate management structure would appear to ensure problematic joint evaluations.

Recruiting the evaluation team

The difficulty of recruiting qualified evaluators was a recurring theme in our interviews, as it was in previous meta-evaluations related to single-agency evaluations. High-quality evaluations require high-quality evaluators. There are too few experienced evaluators available to conduct joint evaluations, let alone real-time joint evaluations. The skills required for joint evaluations and single-agency evaluations are not identical. The joint-evaluation team leader needs to have both management and evaluation skills, and be capable of dealing with the politically sensitive issues that taking a broader perspective requires – in other words, to have a combination of technical, political and inter-personal skills. Team members need to look beyond the single-agency issues to which they are accustomed, and take a sector-wide approach. Having said this, the results on joint-evaluation quality

(discussed in Section 3.4 below) demonstrate that agencies are accessing good-quality evaluators. However, respondents stressed repeatedly the constraints of the current market situation where there is competition for these evaluators.

The TEC lesson-learning exercise in February 2006 commented (p 2):

> staffing evaluation teams with consultants remains a difficult challenge. Often the selection of consultants for evaluation teams is based on who is available, rather than who might be best for the job. Senior consultants have enormous amounts of experience, but are booked in advance, often in predictable patterns around the calendar year. It is recommended to select consultants early.[24]

Three respondents noted that the joint-evaluation focus could depend on the team leader's interests. In the IHE case there was a tension within the core group directing these joint evaluations, between those who wanted the IHE to be policy-oriented, and those who wanted it to be more operationally or programme-focused. What actually happened depended on the profile and inclination of the evaluators. For the 2007 IASC Pakistan RTE a decision was made to appoint a team leader who could deal with the political sensitivities of the response, rather than someone with an evaluation background. UN respondents also noted that finding evaluators with the skills to assess humanitarian reform has been problematic.

The make-up of the evaluation team in terms of gender balance, geographical and institutional representation, and sectoral expertise, was another area on the minds of respondents. One noted that where agency personnel are brought in from outside the affected country to be part of the evaluation team, it is important that they are sufficiently senior and experienced; otherwise they are 'excess baggage'. Several respondents commented on the importance of having a national consultant on the team, to increase contextual understanding and national-level buy-in. In our joint-evaluation sample, 11 evaluations were carried out by mixed (international/national) teams, 6 by international teams, and 1 by a national evaluator. Using a rough calculation,[25] we can conclude that the quality of the mixed-team evaluations was higher. This corroborates findings from earlier meta-evaluations which indicating that a mixed team brought additional skills that may not be found in international teams.

The TEC (2006) lessons-learning exercise concluded:

> In the context **the number of experts from Asia** on the various TEC evaluation teams would also be a matter of concern since many of the affected countries have high levels of expertise. Multi-agency evaluations can be an opportunity to develop capacities in evaluation and development research, but it requires proactive invitations to institutions, and academics from the affected regions to participate.

The TEC capacities evaluation included a breakdown of expenses (Annex 8), and appears to be representative of the TEC reliance on international consultants. Of the total cost of some US$393,000, approximately 1.4 per cent was disbursed on national consultant fees, with 70 per cent disbursed on international consultant fees and travel.

Process and partnership

One of the most important benefits of joint evaluations emphasised by respondents was the way in which they support the building of trust and social capital – a classic case of process being as important as product. As the ECB joint-evaluation guide comments: 'it's important to recognise that you are managing not just an evaluation but also a collaboration' (ECB, 2007, p 12). Similarly, the IHE joint evaluation guide notes: 'decision-making is as much a political process as it is a technical one' (IHE, 2007, p 3). ECB respondents reported that joint evaluations have led to strengthened partnerships in programming, for example in Niger and Java. Country offices reported that joint evaluations have opened up space for increased collaboration between agencies.

All respondents noted that relationship-building is an aspect of joint evaluations that is difficult to measure but should not be ignored. As one said, where the evaluation process itself works well, the uptake of findings is usually good. ECB agencies also stressed the area of mutual accountability. This led to a focus on consultation with the affected population (see Section 3.4.1), and greater attention to recommendations.

The 2007 survey of those involved in the TEC also noted, 'Particularly appreciated was the transparency of the process and the strong focus on information sharing that was put in place from the early stages. [It was a] very inclusive, open coalition,

allowing at various levels and throughout the process for wide consultation' (p 2). This is perhaps reflected in the finding that 80 per cent of respondents felt that they were able to input into the TEC process at times when they wanted to.[26]

3.3.3 Involvement of government and agency stakeholders

In this section we discuss our second two hypotheses: that joint evaluations tend to be driven from the centre (ie, headquarters) rather than from the field, and that the involvement of national government in joint evaluations is weak. We found involvement of the national offices of international agencies in some cases, but less inclusion of national governments and other national institutions.

In-country agency buy-in

All the joint evaluations in our sample were HQ-initiated – and perhaps this is not surprising given their complexity and politically sensitive nature, and the often limited evaluation capacity at country level. Samoff (2005) has emphasised the importance of linking centrally initiated evaluations to locally expressed demand and ensuring that national/local reference groups have real authority, including a direct role in reporting and approval. Otherwise, joint evaluations may receive unenthusiastic cooperation from the field. Several of our joint evaluations did not establish sufficient rapport with country offices at the outset, which made the process problematic and hindered follow-up to recommendations. Interviewees stressed the importance of making clear the added value of a joint evaluation to country offices, and involving them in the development of ToR.

The case of the two IASC RTEs in Pakistan is instructive. In the first RTE in 2006, after the Kashmir earthquake, the evaluation team arrived without prior in-country preparation. In the case of the second RTE in 2007, after flooding and cyclone Yemyin, the evaluation manager went to Pakistan to negotiate the ToR and set up what turned out to be an effective local steering committee. While the evaluation process in the latter case was far from smooth, partly because of political sensitivities, the inclusive approach paved the way for greater agency buy-in within Pakistan. In further contrast, the in-country steering committee set up to guide the IASC Mozambique RTE was less successful, because its ToR were not clear.

The ECB guide to joint evaluation notes the differing cases of the Guatemala and Indonesia joint evaluations:

> The idea and the objectives for the ECB-supported Guatemala evaluation came from headquarters. The team in Guatemala felt that this was another HQ-driven initiative, so their participation in steering committee meetings was reluctant or non-existent. Turnover was high. The agencies on the ground tried to customize the objectives, but in retrospect believe it would have been better if they had started from scratch. This negatively impacted the evaluation process and thus the usage of the findings. In contrast, the idea for the ECB-supported joint evaluation in Jogyakarta also came from headquarters. However, the participating agencies on the ground took the lead on defining their objectives, with advice from headquarters. This helped ensure the partners were more engaged and in control of the evaluation process. (ECB, 2007, p 10)

There appears to be a pattern emerging: joint-evaluation experiments with a longer-term perspective, in particular ECB and IHE, have learnt from earlier experiences and realised the importance of facilitating in-country buy-in. The ECB Indonesia evaluation built on the earlier ECB experience in this respect, as did the second IASC evaluation in Pakistan. The challenge is not to lose this learning in future joint evaluations.

National government buy-in

Our hypothesis that involvement of the national government is weak in joint evaluations was found to hold. Government participation and ownership was problematic, and identified by several respondents as a future key challenge for joint evaluations. The constraints to achieving such buy-in should not be underestimated. In general, involvement in joint evaluations of government and national institutions may be no better than for single-agency evaluations. This was despite many of the evaluations being carried out in countries with reasonable or good government capacity – in particular India, Sri Lanka and Indonesia. It was also despite the fact that our sample mainly deals with natural disasters, where issues of government independence and neutrality are likely to be less acute than in conflict-related situations – although the conflicts and political sensitivities around areas affected by natural disasters in Pakistan, Sri Lanka and Indonesia should be noted. How far the

government is involved will depend partly on the level of joint programming; however, several of the joint evaluations in our sample reviewed government as well as agency performance. The main form of government interaction would appear to be through debriefings on evaluation results. Paradoxically, some of the few examples of inclusion of government were in conflict-related joint evaluations – the IHEs in Chad and Liberia.

The OECD-DAC guide to managing joint evaluations (DAC, 2006) notes that one of the reasons for carrying out a joint evaluation in the development context is to reduce transaction costs for national governments, a point also made more forcefully by Feinstein and Ingram (2003). Humanitarian joint evaluations seem to take their impetus more from the needs of humanitarian agencies, perhaps reflecting differences in aid modalities.

Respondents noted some of the constraints to establishing host-country buy-in: suspicion and lack of trust between international agencies and governments; the perceived need for independent reviews; lack of government capacity in particular in complex emergencies; and the focus of much EHA on international agency performance. There is a trade-off here between independence and government participation to promote utilisation. One interviewee commented, in relation to the TEC:

- After the evaluation got underway, it was recognised that there should have been greater country engagement. But this would have taken considerable time to achieve. For example TRIAMS achieved country level buy-in, but it has taken 2.5 years to reach that point.

- There is a question mark as to how open affected governments would have been to the TEC. The desire to have governments involved was a spin-off from the Paris Declaration on Aid Effectiveness, but it was still too early to make this a reality.

DAC (2005a) suggests that national-government involvement has also been problematic in joint evaluations in the area of development, and that there is a:

- strong feeling of frustration with the present state of affairs, especially as regards the level of partner-country participation in evaluation work

- growing awareness among developing-country representatives of the need for them to play a proactive role in setting and implementing the evaluation agenda in their countries

- clear understanding of the opportunities and benefits, as well as of the problems and challenges, of evaluation work and of carrying it out jointly with donors

- clear interest in learning from evaluation models and success stories in other developing countries.

In contrast, a 2006 single-agency WFP evaluation of its country programme in India (WFP, 2007) included an evaluation team member from the Indian planning body, the National Planning Commission, as well as the former head of the National Planning Commission, in order to promote government buy-in, participation and utilisation, while using an international consultant as the evaluation team leader. The Indian mission in Rome also attended pre- and post-evaluation briefings. This level of national involvement is unusual, however.

The impression from our review is that agencies proposing joint evaluations could be more creative in attempts to involve government and local institutions, especially for natural disasters, whether this is on steering committees, as peer reviewers, or as evaluation team members. Having said that, we recognise many of the difficulties in involving governments where there may be limited capacity, or where joint programming has not taken place. This suggests that small, manageable initiatives, such as briefings for governments, including pre-evaluation briefings on focus, methodology and evaluation-team composition, would be a sensible start. In addition, if joint evaluations are directed conceptually by utilisation-focused approaches, there is greater likelihood of more government engagement.[27]

3.4 Evaluation quality

One of our meta-evaluation hypotheses was that the overall quality of joint evaluations would be higher than that of single-agency evaluations. In particular, we

anticipated that methodologies would be more rigorous, and the crosscutting themes of gender, protection and advocacy would be better covered. We hypothesised this because, in comparison to single-agency evaluations, joint evaluations are likely to: examine broader policy issues; allow access to a wider pool of evaluators; draw on multiple-agency evaluation experience; and be better resourced. In this section, we examine six areas of the Pro Forma in which longitudinal analysis was possible, to determine whether our hypotheses hold, and integrate this analysis with findings from agency and evaluator interviews. Sections 3.5 and 3.6 below cover other areas of the Pro Forma. For the sake of brevity, we have not included all areas of the Pro Forma in the analysis below, and the results on areas not covered in detail here can be found in Annexe 3.6.

3.4.1 Primary-stakeholder consultation and participation

We hypothesised that there is greater opportunity for primary stakeholders (ie, the affected population) to be consulted/surveyed in joint evaluations than in single-agency evaluations. As can be seen from Table 3.3, the performance of joint evaluations in this respect is considerably better than that of single-agency evaluations. Even so, only 47 per cent of evaluations were rated satisfactory or good in this area. In contrast, 80 per cent of evaluations were rated as satisfactory or good in terms of consultation with key stakeholders such as agency and government staff. As noted in previous meta-evaluations, this low rating is partly a result of some evaluations not making transparent the scope of affected-population consultation where this had clearly taken place.

Table 3.3 Pro Forma area 2.4, Consultation with and participation by primary stakeholders (%)

Rating	2001–04 *aggregate*	2006/07
Good	8	31
Satisfactory	19	16
Unsatisfactory	39	37
Poor	34	16

As detailed above (Section 3.1.3), evaluations from the UN system made up 33 per cent of the meta-evaluation sample, with 36 per cent from NGOs. NGO evaluations

achieved better ratings in this Pro Forma area than did UN evaluations, with 58 per cent rated as good and 8 per cent as satisfactory for the former, as opposed to 20 per cent rated good and 10 per cent as satisfactory for the latter.

During interviews respondents consistently noted the importance of proactive attention to consulting with the affected population. One respondent commented:

> Strengthening downwards accountability has been key. The way the ECB evaluations have been structured was important to achieve this. In each ECB evaluation, the team leader was asked to focus particularly on ensuring that primary stakeholders are heard. The extent to which this was achieved partly depended on the skills of the team leader. ECB HQ staff kept pushing that focus, with country programmes, in the ToR and with the evaluators. It required someone senior to set it as a priority and to maintain that focus.

Details of five evaluations that noted the numbers of the affected population consulted are presented in Table 3.4. This wide range suggests that evaluators are usually left to themselves concerning the scale of consultation. Attempts are made to take a roughly representative and random sample; however, no statistical analysis is included in any joint evaluation reviewed, and there does not appear to be any clear rationale as to the numbers consulted, ie consideration of the advantages of consulting with 500 as opposed to 400 people. Further guidance in this area is needed.

Table 3.4 Numbers of affected population consulted

Evaluation	Number of affected population consulted
ECB Guatemala	124
ECB Java	318
IASC Mozambique	400
DEC tsunami	500
TEC capacities	2,055

Consultation in our sample takes a standard approach – usually one-time focus-group discussions supplemented by individual interviews – but the balance chosen between these two methods is rarely explained. An opportunity was lost with the TEC to carry out more rigorous surveys focusing on results from the perspective of the affected population, which would have required an intensive longitudinal

approach.[28] This is not usually feasible in single-agency evaluations, where setting up quasi-research projects can rarely be justified. In the case of the TEC, however, it would have been appropriate and feasible – given the level of resources and regional social-science expertise available, as well as funding in excess of US$13 billion for the response. Instead the TEC relied on the 'tested but tired' EHA approaches – dispatching small teams of expatriates for short periods of time, although these teams nevertheless made their best attempts at community consultation within the constraints imposed.

The importance of including the affected population in the evaluation process was also recognised, for example in discussion with evaluation managers from the Inter Agency Health Evaluations, who suggested that representatives from the affected population should be included on the evaluation steering committee, but also noted that they had not found an effective way of realising this. Only the ECB evaluation in Indonesia included the affected population as active participants in the evaluation (rather than only as participants in focus-group discussions), and invited them to the discussions of evaluation findings. This is an example of good practice.

3.4.2 Evaluation methods

We hypothesised also that joint evaluations would have more rigorous methodologies than single-agency evaluations, which in turn would lead to higher-quality evaluations. Table 3.5 illustrates that this was the case.[29] Between 2004 and 2006/07, there was no significant difference in the percentage of evaluations rated as satisfactory or better (59 per cent in 2004, and 67 per cent in 2006/07). But the percentage of joint evaluations rated good increased significantly (from 11 per cent in 2004, to 25 per cent in 2006/07).

Table 3.5 Pro Forma area 2.3, Appropriateness of the overall evaluation methods (%)

Rating	2004	2006/07
Good	11	25
Satisfactory	48	42
Unsatisfactory	37	28
Poor	4	5

As Annexe 3.8 illustrates, almost all of the joint evaluations used familiar EHA techniques, varying mixes of document review and interviews with key stakeholders and affected population, usually stating that triangulation would take place, but less often comparing different data sources.

Samoff (2005), commenting on the basic-education joint evaluation, sees joint evaluation as an opportunity to innovate, to go beyond constraints of risk-averse single agencies, for example using new combinations of standard methodologies, He also notes that innovation is likely to drop off as the joint evaluation proceeds, when compromises and lowest-common-denominator solutions start to dominate to keep it on track. There was limited methodological innovation found in our sample. Respondents also noted that data analysis was often problematic, even given the normal constraints around data collection and analysis in the evaluation of humanitarian action: too much data were collected and could not be analysed, or data-analysis skills were weak, or the basis for conclusions was not clear.

In the evaluations rated as good in this area, there was much evidence of good practice, and two evaluations stood out. In the ECB evaluation of tsunami operations in Indonesia and Thailand evaluation, special efforts were made to reach local people inland, who had not necessarily benefited from the tsunami response. In the IASC Mozambique evaluation, despite it being a real-time evaluation:

- constraints and limitations are clearly set out
- the evaluation team accessed 700 relevant documents, which were indexed and provided with a search engine; and a summary of lessons learned from previous evaluations was made
- focus-group meetings with over 400 members of the affected population were held, in 16 sites, using male and female interviewers so men and women could be interviewed separately if needed
- semi-structured interviews were held with a further 95 key informants
- there was triangulation between the different data sources.

We found that a higher-quality methodology fed into a higher-quality evaluation report. The 13 joint evaluations rated as good or satisfactory on quality of methodology achieved a 70 rating on a composite index of evaluation quality, while the 5 joint evaluations rating unsatisfactory and poor achieved a composite index of

49.[30] While this is indicative only, given the small size of the sample, it suggests that improving methods will feed into higher-quality evaluation.

3.4.3 Use of international standards

We hypothesised that joint evaluations are likely to pay greater attention to international standards and guidelines than are single-agency evaluations. The standards joint evaluations were expected to assess interventions against included international humanitarian and human rights law, the Red Cross/NGO Code of Conduct, Sphere, and Guiding Principles on Internal Displacement.

Table 3.6 Pro Forma area 2.5, Use of and adherence to international standards and guidelines (%)

Rating	2002–04 *aggregate*	2006/07
Good	6	47
Satisfactory	14	26
Unsatisfactory	23	18
Poor	57	9

Past ALNAP meta-evaluations have criticised agencies for failure to assess performance against standards to which they are signatories. This is shown in Table 3.6, where 80 per cent of evaluations for 2002–04 rated unsatisfactory or poor in this area, including 57 per cent rated as poor. Joint evaluations of 2006/07 performed considerably better, suggesting that joint evaluations are more likely to pay attention to inter-agency standards.

For the 13 evaluations in the 2006/07 sample that included satisfactory or better attention to standards, 10 included assessment against Sphere, 2 against the Red Cross/NGO Code of Conduct, and 2 jointly used Sphere and the Red Cross/NGO Code of Conduct. All ECB evaluations were rated as good on this Pro Forma area. NGO evaluations achieved considerably better ratings than UN evaluations in this area, with 92 per cent rated as good and 8 per cent rated as satisfactory for the former, and 20 per cent rated as good and 30 per cent rated as satisfactory for the latter. Part of the reason for this result is the consistent use of Sphere standards by NGO evaluation teams.

3.4.4 Use of the DAC criteria

Most evaluation of humanitarian action is organised around the DAC criteria, and agencies have performed relatively well in their use. Use of these criteria has consistently improved during the period covered by ALNAP meta-evaluations to date (Figure 3.2).[31] Table 3.7 illustrates aggregate performance of joint evaluations' use of the DAC criteria, as opposed to single-agency evaluations. Given the improvement of single-agency evaluations over time, an additional column for results for 2004 only has been included in the table.

Figure 3.2 Longitudinal assessment of the use of DAC criteria

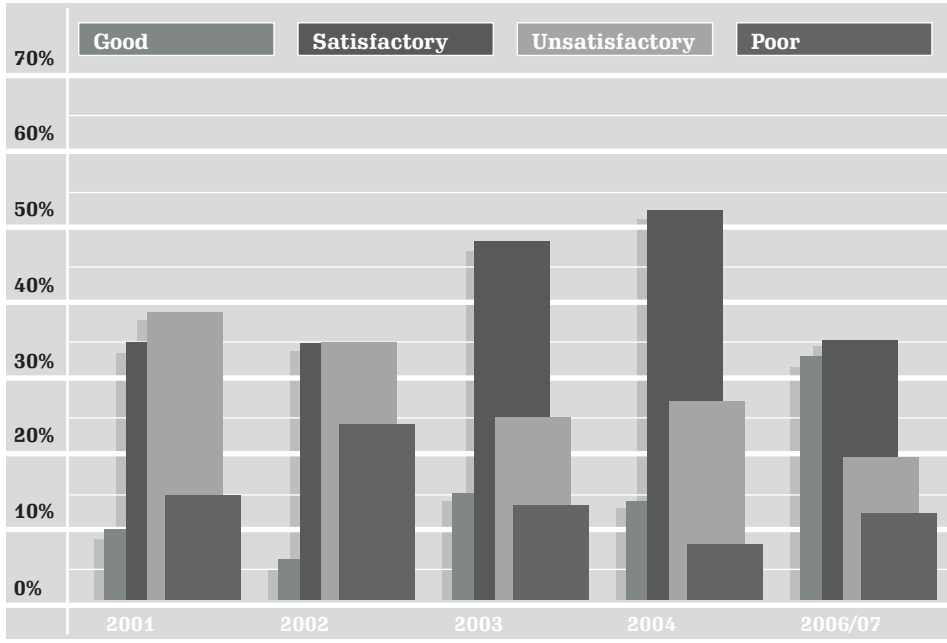


Table 3.7 Pro Forma area 4.3, Application of EHA criteria (%)

Rating	2002–04 *aggregate*	2004	2006/07
Good	11	13	34
Satisfactory	42	53	34
Unsatisfactory	32	26	20
Poor	15	8	12

Compared to the 2002–04 aggregate, joint evaluations can be seen to perform considerably better, but only marginally better compared to the sample in the 2004 meta-evaluation. The main difference is the increased number of evaluations rated good as opposed to satisfactory in 2006/07.[32]

Table 3.8 Pro Forma areas 4.3i-vii, Individual DAC criteria (%)

Efficiency	2002–04 *aggregate*	2006/07
Good	10	20
Satisfactory	34	27
Unsatisfactory	35	23
Poor	21	30
Effectiveness		
Good	12	47
Satisfactory	66	39
Unsatisfactory	20	14
Poor	2	–
Impact		
Good	7	19
Satisfactory	40	62
Unsatisfactory	37	19
Poor	16	–
Relevance/approproateness		
Good	19	47
Satisfactory	56	36
Unsatisfactory	19	11
Poor	5	6
Connectedness	2002–04 *aggregate*	2006/07
Good	16	62
Satisfactory	60	32
Unsatisfactory	18	6
Poor	6	–
Coverage		
Good	12	29
Satisfactory	39	41
Unsatisfactory	34	21
Poor	15	9

Table 3.8 Pro Forma areas 4.3i-vii, Individual DAC criteria (%) *continued*

Coherence		
Good	3	3
Satisfactory	28	9
Unsatisfactory	28	47
Poor	40	41

Disaggregated performance on individual criteria is illustrated in Table 3.8 for 2002–04, and 2006/07. Joint-evaluation performance is higher in all areas except for coherence (discussed further in Section 5.4 below). In 2006/07, there is a higher percentage of good/satisfactory ratings, as well as a higher percentage of evaluations rated as good. Of note is the considerably better performance on the impact and connectedness criteria, suggesting that joint evaluations have been able to analyse longer-term issues and recovery more fully. This may also relate to 70 per cent of our sample evaluating responses to sudden-onset natural disasters, in which the immediate relief phase is often quite short and attention shifts quickly to recovery. Assessment of efficiency is only marginally better for joint evaluations than single-agency evaluations, with the majority of the 2006/07 sample (53 per cent) receiving ratings of unsatisfactory or poor. As noted in previous meta-evaluations, efficiency is one of the least-understood of the DAC criteria, and requires specific evaluation skills not always available.[33]

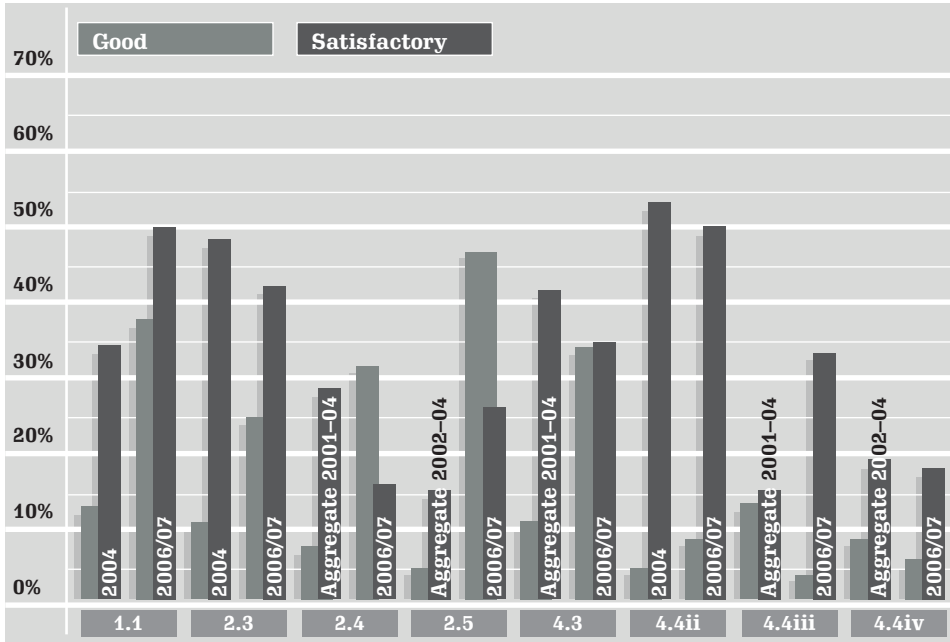
3.4.5 Gender equality

Table 3.9 demonstrates that our hypothesis that joint evaluations are stronger than single-agency evaluations on crosscutting issues such as gender equality did not hold. Results for 2006/07 are similar to those of the 2004 meta-evaluation, with gender analysis rated as satisfactory or good in 59 per cent and 58 per cent of the evaluations, respectively. However, no joint evaluations assessed the intervention against the agency's gender policy. Given that most agencies now have a gender policy, this is a serious gap, and demonstrates that joint evaluations are not more likely to deal with policy issues in this area.

Table 3.9 Pro Forma area 4.4.ii, Gender equality (%)

Rating	2004	2006/07
Good	5	9
Satisfactory	53	50
Unsatisfactory	35	26
Poor	7	15

Figure 3.3 Selected Pro Forma ratings, 2002–04 and 2006/07



Note The numbers on the horizontal axis refers to the Pro Forma sections, as follows:

1.1 Terms of reference
2.3 Appropriateness of the overall evaluation methods
2.4 Consultation with and participation by primary stakeholders
2.5 The use of and adherence to international standards and guidelines
4.3 OECD-DAC criteria – aggregate
4.4.ii Gender equality
4.4.iii Protection
4.4.iv Advocacy

3.4.6 Summary

Figure 3.3 highlights overall performance of joint evaluations in eight Pro Forma areas, in comparison to single-agency evaluations. For the six areas covered in this section, the average proportion of joint evaluations rated good was 27 per cent, as against 9 per cent for single-agency evaluations; and the average percentage for joint evaluations rated satisfactory was 40 per cent, as opposed to 34 per cent for single-agency evaluations. We can conclude that joint evaluations are of considerably higher quality than single-agency evaluations.[34]

Box 3.2 The Tsunami Evaluation Coalition (TEC)

Given the significance of the tsunami evaluation as the first system-wide joint evaluation since the Joint Evaluation of Emergency Assistance to Rwanda, published in 1996, the authors of this meta-evaluation were asked to include a consideration of the TEC process and quality, in terms of strengths and weaknesses. Table 3.10 shows performance of the four TEC reports assessed against select areas in the Pro Forma.

Table 3.10 TEC evaluation reports rated against select areas of the Pro Forma (percentage of evaluations)

Pro forma area	Good	Satisfactory	Unsatisfactory	Poor
2.3 **Appropriateness of the overall evaluation methods**	25	75	–	–
2.4 **Consultation with and participation by primary stakeholders**	17	50	17	16
2.5 **The use of and adherence to international standards and guidelines**	37	63	–	–
4.1.ii **The agency's management and human resources**	63	37	–	–
4.2.i **The needs and livelihoods assessments that informed the intervention**	25	75	–	–
4.3 **Application of EHA criteria (aggregate)**	31	44	19	6
4.4.ii **Gender equality**	25	75	–	–
4.4.iii **Protection**	–	75	12.5	12.5
4.4.iv **Advocacy**	25	25	50	–
5.1.iii **Recommendations**	12.5	12.5	75	–

CONTINUED

Box 3.2 The Tsunami Evaluation Coalition (TEC) *continued*

The TEC evaluations rated highly on almost all sections of the Pro Forma, and were overall of good quality. This is a significant achievement, and credit should go to the many people committed to the TEC. Development of evaluation methods, review of needs and livelihoods assessments, and attention to gender equality, were all particular strengths. The exceptions were the Pro Forma areas of advocacy and recommendations. Greater attention could have been paid to issues of identification and lobbying for marginalized groups, and advocacy around buffer zones.

The authors of this meta-evaluation anticipated clearer direction of recommendations to particular users, and greater specificity, than was found in the TEC recommendations. The TEC reports also repeated recommendations that have been made a number of times in other reports over the last several years. As one respondent noted, repeating past recommendations that have not been followed suggests that the TEC was making the wrong kinds of recommendations, or at least did not adequately explore past constraints to follow-up and thus ensure that the recommendations had a better chance of being implemented.

Interview findings

The TEC process was complex, and at times problematic – this is perhaps not surprising given the magnitude of the disaster, the level of funding and number of agencies involved, and that this was the first joint evaluation the system had attempted in ten years. In interviews, a number of issues were raised which corroborate findings from the TEC After Action Reviews.

The TEC Secretariat coordinated the evaluation effectively. Using ALNAP to host the Secretariat allowed fast start-up, and initial seed funding. The Secretariat was effective in keeping a wide range of stakeholders informed. It also managed part of the funding to ensure successful completion and launch of reports. Because of this effective coordination there is now a budget line in ALNAP for starting-up system-wide joint evaluations if required in future.

CONTINUED

Box 3.2 The Tsunami Evaluation Coalition (TEC) *continued*

Several respondents noted the **lack of an overall conceptual framework** as a core issue that impacted negatively throughout the evaluation. Without this, the TEC faced some of the constraints that have challenged much EHA, whether joint or single-agency evaluation – for example lack of attention to utilisation, and failure to engage adequately with national capacity. As one respondent noted:

> The idea of working to a common framework fell through. Partly, as a result, we ended up with a thematic model. But it didn't really work. In retrospect we should have had an output-/impact-oriented approach, with a number of different evaluations in different countries using a common framework for comparative purposes.

Roles and responsibilities could have been more clearly defined. Roles and responsibilities of the Core Management Group (CMG), evaluation team leader and evaluation team members were inadequately defined, leading to some confusion and tension during the TEC process. Different agencies were delegated to lead the thematic evaluations, but in the absence of a common framework and with inadequate coordination, negatively impacting on the jointness and coherence of the whole exercise. Respondents noted that the CMG and TEC structure was symbolic of a **top-down and centralised approach** that dominates the sector. 'It was a very northern-based initiative, that we did not do enough to mitigate', one respondent noted.

A TEC After Action Review[35] noted the **constraints faced by the TEC in tying in to national capacity**, thereby mirroring some of the problems of the response. Although efforts were made to include regional actors, especially on Peer Review Groups and as key panel members at all the launches, a number of respondents felt that more could have been done to involve governments in the region and to have given them more of a voice in responding to the reports. This is a clear lesson emerging from the TEC.

Lost opportunities to consult with the affected population: as noted in more detail in Section 3.4.1, the TEC missed the opportunity to develop longitudinal assessments which would have supported downward accountability, for example through repeat surveys or focus groups over the period of a year.

CONTINUED

Box 3.2 The Tsunami Evaluation Coalition (TEC) *continued*

Limited assessment of impact and policy. One participant in a TEC lessons-learning exercise in February 2006 (TEC, 2006, p 2) noted:

> there was insufficient consideration of impact in the TEC…. Process (coordination, needs assessment, funding, etc.) is important, but what matters is outcomes… I have never come across an evaluation where so much has been spent, so much discussed, and so many conclusions drawn with so little reference to, or evidence on, impact and outcomes.

The TEC did not achieve its planned policy focus. As one respondent noted: 'Despite a desire to be so, in the end the TEC wasn't that policy-focused. Evaluators write evaluations. If we wanted policy-focused outputs we should have gone to policy analysts.'

Recommendations and follow-up: TEC recommendations were of mixed quality in different reports, with some at a very general level (eg in the LRRD report) and others more clearly targeted (eg in the synthesis). Although follow-up was poorly planned at the outset, this was addressed later on. There were two international launches, a considerable amount of media work immediately after the release of the reports, and a part of ALNAP's biannual meeting in Rome in December 2006 was dedicated to discussing the TEC findings and implications for the system. With a longer-term focus, an OCHA-led Tsunami Advisory Group was set up to follow up on the findings and recommendations.

3.5 Do joint evaluations have a broader policy focus?

3.5.1 Introduction

One of the frequently cited benefits of joint evaluations, and therefore one of the reasons for doing them, is that they are able to look at the big picture and evaluate collective action within the wider context (eg ECB, 2007). They can also tackle questions that cannot be addressed by any one agency, for example on coordination and coherence of the response: how agencies relate to each other and also to government authorities (DAC, 2006). Thus, joint evaluations should be well-placed to engage with and to influence policy. This is captured in two of our working hypotheses, that joint evaluations are more likely than single-agency evaluations to address policy issues (as well as programme performance), and that joint evaluations situate their findings within wider debates within the humanitarian sector. Referring back to the typology in Table 3.1, the level at which the evaluation engages with policy depends upon the scope of the evaluation and which actors are involved. In a multi-sectoral joint evaluation carried out by 'like-minded agencies', it is most likely to influence the individual and collective policies of that group of actors, but in a system-wide multi-sectoral joint evaluation, one would expect higher-level engagement with policies that affect the whole international humanitarian system.

This section explores the extent to which joint evaluations in our sample have engaged with policy, first by reviewing the terms of reference of the evaluations and then looking at three areas of enquiry from the Quality Pro Forma. Those three areas are: the quality of the contextual analysis (important for placing the evaluation findings in the bigger picture); the extent to which coordination and coherence have been evaluated as part of the DAC criteria; and how well two of the crosscutting issues, protection and advocacy, have been addressed.

3.5.2 What do the commissioning agencies want?

To what extent do commissioning agencies actually want a bigger-picture/policy focus, and how explicit is this in the ToR? The ToR for the IASC RTEs are among the

most policy-oriented of the set, making direct reference to the UN-led humanitarian reform process. OCHA interviewees for the meta-evaluation confirmed that these exercises had really helped the agencies involved to see the big picture in terms of how everyone responded (not just the UN country team but also NGOs and other agencies). One observed, 'We were forced to step back and take a broader view and look at humanitarian reform, in particular coordination'. The IHE initiative is also explicit in terms of its focus on policy-level analysis, demonstrating clearly how this differs from single-project evaluations and how policy-evaluation techniques must be used in sector-wide evaluations, although still combined with some project-level evaluation techniques (IHE, 2007).

Despite the disparate ToR for the TEC studies, contributing to improved policy is clear in most of them. But the outcome has been disappointing in terms of meeting these expectations, and the TEC has engaged less at policy level than anticipated, for example rather little with the UN humanitarian reform process (see also Box 3.2 above).

In contrast, the ECB ToR have tended to be more operational and programme-focused, although most do refer explicitly to the importance of evaluating coordination and coherence between agencies. Almost all those interviewed for this meta-evaluation from the ECB NGOs stressed the benefit of 'stepping back from the operational detail of individual agencies to see the bigger picture', in the words of one ECB advisor. Interestingly, at least three agency staff interviewed for this meta-evaluation have commented that, no matter what the ToR say about a policy focus, what happens in practice depends upon the interest and bias of the evaluation team selected, and especially the team leader.

3.5.3 Contextual analysis

Table 3.11 Pro Forma area 3.1, Contextual analysis (%)

Rating	2004	2006/07
Good	20	47
Satisfactory	45	29
Unsatisfactory	28	21
Poor	7	3

Good contextual analysis is important for any evaluation, but it is essential for a policy-focused evaluation in order to draw policy-level conclusions. The Pro Forma results show that contextual analysis is stronger for joint evaluations – almost half rate as 'good' – than for the sample of 30 evaluations in 2004 (Table 3.11). The benefits of a good contextual analysis were particularly apparent in the ECB Niger evaluation. This provided a very strong analysis of what had been described by agencies and the media as 'famine' in Niger, paying attention to the economic and historic determinants of the crisis and challenging agency perceptions of famine as well as how the international response may have exacerbated the crisis.

A number of joint-evaluation reports have included chronologies, usually in an annexe. Two of the ECB reports are models of good practice: the Thailand and Indonesia tsunami evaluation, and the Niger evaluation. The format is simple but effective in each, plotting key events (including political events) that have influenced the crisis against the timings and process of the humanitarian response. Yet one evaluator interviewed for this meta-evaluation described how they had to fight to include the chronology, and especially the political analysis, albeit with an agency that was at that time less familiar with joint evaluations.

3.5.4 Coordination and coherence

Almost all of the joint evaluations paid attention to coordination: 84 per cent rated good against this area of enquiry in the Pro Forma and 9 per cent rated satisfactory – a considerable achievement. There was some difference in how well coordination was covered according to the origin of the report: reports commissioned by UN agencies were strongest as a group on this topic. There are a number of examples of good practice, including the ECB evaluation of the tsunami response in India and Sri Lanka, unusual because it includes an in-depth analysis of the role of government in coordination.

Surprisingly, this pattern is not repeated for 'coherence'. As Table 3.12 shows, this year's sample of joint evaluations performs substantially worse against this criterion than did the 2002–05 aggregate. Only one evaluation from the 2006/07 sample, the TEC LRRD evaluation, was assessed as 'good' in this area because it included a comparative analysis of government and international-agency LRRD policies.

Table 3.12 Pro Forma area 4.3.vii, Coherence (%)

Rating	2002–05 *aggregate*	2006/07
Good	3	3
Satisfactory	28	9
Unsatisfactory	28	47
Poor	40	41

Interpreting coherence is a problem that has been highlighted in previous meta-evaluations (Wiles, 2005, p 159). In the ALNAP EHA guide it is defined as: 'the need to assess security, developmental, trade and military policies, as well as humanitarian policies, to ensure that there is consistency and, in particular, that all policies take into account humanitarian and human-rights considerations' (Beck, 2006, p 21). While this may be relevant to humanitarian action by donor governments, it is hard to apply to NGO or UN humanitarian programmes. For the purposes of this meta-evaluation, we interpreted coherence as policy coherence between international humanitarian agencies, and coherence with national government policy.

In practice, coordination and coherence were frequently dealt with in the same section in evaluation reports; this usually meant that coherence was overlooked and the entire section was devoted to coordination. How far the international agencies' response was coherent with the policies of the national government was an unexpected gap in most reports, especially for some of the tsunami evaluations (with the one exception noted above). Analysing the extent to which agency policies were coherent with one another was a glaring gap, surprisingly so when the work of a number of agencies was being evaluated together.

3.5.5 Policy-related crosscutting issues: protection and advocacy[36]

As explained above (at the beginning of Section 3.4), we had expected the joint evaluations to be strong on crosscutting issues. However, we found that this was not the case with either protection or advocacy, as demonstrated in Tables 3.13 and 3.14. It is striking that the majority of joint evaluations were unsatisfactory or poor in their coverage of protection, and that so few rated 'good', especially as this has been highlighted as part of humanitarian action in recent years.

Where protection issues were raised, it was often in relation to women and sexual and gender-based violence (SGBV), or in relation to land rights in some of the tsunami-related evaluations. One of the reasons why protection is so poorly covered may be to do with the skewing of the sample towards natural disasters: protection may be seen as a low priority in these contexts. But as some of the more policy-oriented tsunami evaluations reminded us, natural disasters do not occur in isolation from political context. Protecting land rights was a relevant issue, and some of the areas worst affected by the tsunami had also suffered years of conflict (notably, Aceh and parts of Sri Lanka). No examples of good practice have emerged in the area of protection.

Table 3.13 Pro Forma area, 4.4.iii, Protection (%)

Rating	2001–04 *aggregate*	2006/07
Good	14	3
Satisfactory	15	33
Unsatisfactory	22	20
Poor	48	43

Only one report in the whole sample has a section dedicated to advocacy – the TEC evaluation report on coordination. Some others mention advocacy, but the majority ignore it completely or make passing mention. Once again, is this because most evaluations were to do with natural disasters, and advocacy is regarded as less relevant here than in conflict-related humanitarian crises? But if protection issues are pertinent in so-called natural disasters, then advocacy must be as well: for example, through lobbying duty-bearers (often governments) to protect the rights of the vulnerable who may be subject to abuse or exploitation. It is notable that the TEC synthesis report was rated as unsatisfactory for coverage of both protection and advocacy.

Table 3.14 Pro Forma area 4.4.iv, Advocacy (%)

Rating	2002–04 *aggregate*	2006/07
Good	8	6
Satisfactory	20	19
Unsatisfactory	40	31
Poor	32	44

3.5.6 Summary

The joint evaluations have generally done well in placing their findings in the wider context, in painting the bigger picture. But their engagement with policy issues and their ability to move policy debates forward are not as strong as expected. Although one respondent commented, 'JEs are the best way of figuring out what is happening in the sector as a whole', this potential does not seem to have been fully realised. The IASC RTEs have perhaps made greatest progress in this respect, in relation to the UN-led humanitarian reform process. The health sector joint evaluations carried out by the IHE have also engaged at the policy level.

The TEC evaluations, however, have a mixed record. A number of respondents were disappointed that they were not more policy-focused, drawing comparisons with the last system-wide Rwanda evaluation, which did have a strong policy focus even if the impact on improved practice has been less positive.[37] Time will tell if the TEC is able to have a greater impact on improved practice than the Rwanda joint evaluation. On the crosscutting issues of protection and advocacy, joint evaluations have a long way to go and do not seem to be of superior quality to other evaluations of humanitarian action. One respondent suggested that these issues should be specifically mentioned in the ToR to ensure that they are addressed; this had rarely been the case in the joint evaluations reviewed.

3.6 Follow-up and utilisation

3.6.1 Introduction

ALNAP's review of the utilisation of evaluations of humanitarian action opens:

> 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. (Sandison, 2006, p 90)

There is an urgent need to pay attention to utilisation. How can this be done better and what can be learned from examples of good practice?

More specifically, how do joint evaluations fare in terms of follow-up and utilisation? This section goes as far as possible in answering that question. It assesses the quality of evaluation reports in terms of overall accessibility for their intended users, the clarity and logic of their conclusions, and the quality of their recommendations. It then reviews follow-up mechanisms to joint evaluations, most of which we have heard about in agency interviews. However, we stop short of assessing the impact. Not only would this require a separate investigation, but also in many cases it is too early to assess the impact of evaluations that have been completed so recently, including the TEC evaluations.

3.6.2 Evaluation reports: accessibility, conclusions and recommendations

The quality of joint evaluation reports in terms of accessibility is markedly higher than as found in the results of previous meta-evaluations (Table 3.15). Over half of the 2006/07 sample rate as good, compared with only 12 per cent from previous years. This is an encouraging result as the readership of joint evaluations is, by definition, much larger.

Authors of joint-evaluation reports are often under pressure to produce concise reports, yet they also often have to distil much larger amounts of material, both from documentation reviews and from fieldwork, than in single-agency evaluations. The way that many have dealt with this is to have quite detailed annexes and to write the main report more as a synthesis. Our sample reports demonstrate a number of different approaches: the ECB Yogyakarta earthquake report presents its findings per agency in the annexes; the TEC capacities evaluation presents its findings per country in the annexes; a number of others pick out particularly pertinent issues and address them at greater length in the annexes (eg land rights in the case of the ECB Thailand and Indonesia tsunami evaluation; SGBV in the case of the IHE Liberia evaluation). The IASC's RTE in Mozambique departs from the usual presentation of

pages of text, to include photographs, maps and 'call-outs' in sidebars, all of which help to improve its readability.

The reasons for such readable and accessible joint evaluation reports may be partly to do with a higher skill set demanded of evaluation team leaders in joint evaluations. But for some of the higher-profile joint evaluations, the commissioning agencies have employed the services of a professional editor. This was the case with the TEC.

Table 3.15 Pro Forma area 5.2.ii, Accessibility of the report (%)

Rating	2001–04 *aggregate*	2006/07
Good	12	58
Satisfactory	35	39
Unsatisfactory	17	3
Poor	37	0

Overall, the quality of the conclusions in joint-evaluation reports is slightly stronger than for evaluations in previous years, but they are mostly 'satisfactory' rather than 'good' (Table 3.16). We noted an interesting contrast between the RTEs and ex-post evaluations: two of the four RTEs had no final concluding section, and a third presented very brief conclusions without any supporting evidence. The speed at which these RTEs have to be completed may be the reason for this, but the lack of a conclusion does weaken the accessibility of a report.

Table 3.16 Pro Forma area 5.1.ii, Quality of conclusions

Rating	2001–04 *aggregate*	2006/07
Good	21	15
Satisfactory	46	68
Unsatisfactory	22	12
Poor	10	6

There is relatively little difference in terms of overall quality of recommendations between the joint evaluations and earlier meta-evaluations, although a quarter of the joint-evaluation reports scored 'good' in terms of recommendations. As inadequately targeted recommendations are one of the criticisms of joint

evaluations,[38] this is an encouraging result. The strongest sets of recommendations were those targeted to individual agencies or where responsibility was clearly indicated, for example in the ECB India and Sri Lanka tsunami evaluation. Developing clear, specific and targeted recommendations is most challenging for a system-wide joint evaluation, and the TEC reports had a mixed record in achieving this. Not surprisingly, the RTEs reviewed produced some of the most specific recommendations and some indicated where responsibility should lie. But they, too, suffered from a common problem of producing too many recommendations, which can dilute their overall impact: the IASC RTE in Mozambique contains over 30 recommendations. Agency interviewees confirmed that too many recommendations can be overwhelming, and may just 'disappear'.

Where recommendations in the joint evaluations reviewed were weak, this was primarily because they were pitched at too general a level without clearly indicating responsibility, or because they were too numerous and poorly organised.[39] Few recommendations indicated any timeframe or prioritisation for implementation. These are well-known problems that have been recorded in previous meta-evaluations (for example, Wiles, 2005), implying that this message is not reaching evaluators, or they are not heeding it.

Table 3.17 Pro Forma area 5.1.iii, Quality of recommendations

Rating	2001–04 *aggregate*	2006/07
Good	8	26
Satisfactory	42	32
Unsatisfactory	29	38
Poor	22	3

3.6.3 Follow-up mechanisms

The literature on utilisation stresses that how an evaluation is going to be used must be addressed from the outset, not least in terms of identifying and involving the target audience and agreeing a budget for follow-up and dissemination. This is more challenging for joint evaluations than for single-agency evaluations because of the larger range of stakeholders with different needs, creating a demanding burden in

terms of coordination and negotiation. (The improved results of joint evaluations presented in Table 3.17 should be read in this light, meaning that the achievement involved is considerable.)

Even on a relatively small scale, the varying needs of different stakeholders are evident from the DEC's experience. While chief executives of DEC member agencies are most concerned with information that helps to meet public-accountability requirements and with issues relating to their agencies' public profile, the humanitarian directors of the same agencies are interested in a deeper operational analysis that helps them locate their agency's response in the bigger picture, and enables learning from peers. How much greater is the challenge of adopting a utilisation focus in the case of a system-wide evaluation like the TEC, especially with the inclusion of a wider stakeholders such as national NGOs and governments of affected countries? By the admission of members of the TEC Core Management Group, a utilisation focus was overlooked at the outset of the TEC, in the rush to set it up and get buy-in. Although this was picked up half-way through the TEC process (and, encouragingly, this was the best-attended TEC meeting), in the words of one respondent: 'we chronically underestimated the need for follow-up'.

Being utilisation-focused directly relates to country buy-in. As one of the ECB agency respondents put it: 'the best use of an evaluation is made where the country team really has ownership…. And where the process for the evaluation itself has worked well, the uptake of findings is usually good'. The ECB Yogyakarta earthquake evaluation set out to be consciously utilisation-focused from the outset and offers an interesting example of good practice (Box 3.3).

Box 3.3 ECB, Yogyakarta: a positive example of being utilisation-focused

Key to the process was the decision by CRS, the lead agency, to adopt a utilisation-focused approach. CRS at all levels – headquarters, national and local – willingly endorsed this process, encouraging a high level of engagement by other participating NGOs. Although generic ToR were used, the agencies in-country were asked to be really specific about what they wanted from the evaluation, and particularly from the 'big' questions, *before* the evaluation began. Thus, dialogue between the evaluation team leader and in-country steering committee started before the team leader arrived in country.

CONTINUED

Box 3.3 ECB, Yogyakarta: a positive example of being utilisation-focused *continued*

On arrival in country, the evaluation team developed its methodology with the steering committee, in particular discussing the most appropriate interpretation of the DAC criteria. ('Connectedness' thus evolved into issues about recovery.) Once stakeholder consultations and document review were completed, the evaluation team presented summary findings to the steering committee and field staff from the four agencies. Participants worked with evaluation team members to draw conclusions and recommendations. Participants were encouraged to prioritise only a few recommendations, to increase the likelihood of implementation.

At the end of the evaluation, a final meeting was held with the evaluation team, steering committee members, government officers and some local people. The steering committee members presented the initial conclusions, and asked participants to review and amend them and make a few recommendations. This process generated a richer set of conclusions and recommendations with a much greater sense of ownership.

The DAC study on joint evaluations identifies follow-up as a 'weak link in the chain… and suggests that there is no common agreement yet on the kind of follow-up suitable and appropriate for JEs' (DAC, 2005a, p 67). It notes that bilateral donors tend to treat follow-up as they would for a single-agency evaluation, in other words with a management response. This pattern was repeated by some agencies participating in joint evaluations in our sample, for example CARE responds to the ECB joint evaluations with a management response in country, just as they would for a single-agency evaluation. The value of this is that it helps to avoid the pitfall identified by one respondent: 'the recommendations in a JE are often generic and not specific to any one agency. Agencies can end up picking the ones they like, but in single-agency evaluations they do not have this option.'

A couple of respondents gave examples of how critical findings from a joint evaluation can provoke a more defensive response than in a single-agency evaluation, often accompanied by a discrediting of the evaluation team and/or their methodology, especially if the agency concerned played little or no role in selecting the team.[40] At worst this can result in disengagement and no chance of

recommendations being taken up. This response is less feasible to a single-agency evaluation where the team was directly hired by the agency and, in the words of one interviewee, 'there is nowhere to hide'.

One of the more impressive examples of follow-up from our set of joint evaluations comes from the IHE initiative. With learning from early joint-evaluation experiences, the importance of buy-in at country level quickly became apparent, as did how to achieve this through an in-country steering committee and a pre-visit by members of the IHE core working group to help set up the evaluation.[41] A later innovation was to schedule an action-planning workshop, in country and ideally less than a year after the evaluation was completed. The steering group led this workshop, to discuss follow-up with all stakeholders; the evaluators returned for the workshop.

The IHE experience is a reminder that follow-up is usually easier and more likely to happen when the joint evaluation is part of a wider institutional framework. This has been the case with Health Action in Crisis, of which the ECHO/WHO/DFID RTE was a part. Its steering committee meets every six months, providing a natural venue for discussion of the recommendations that have come out of the (now nine) different RTE exercises. A similar process appears to be developing slowly around the IASC RTEs, where recommendations are increasingly discussed in different IASC fora. The UN's Emergency Relief Coordinator Office is apparently monitoring implementation of the Mozambique RTE recommendations on a three-monthly basis.

This kind of institutional follow-up has not yet happened regularly in the ECB project, although one of the ECB agency respondents thought this could be the next step in developing peer accountability. But the ECB joint evaluations do demonstrate interesting examples of positive outcomes from the joint process of the evaluation. One of the most powerful examples comes from Niger:

> Having gained a common understanding of the Niger crisis and established working relationships the partner agencies continued to meet long after the evaluation team had left Niger. They formed an NGO coordination forum, called the GDCI and invited other agencies to join… They… successfully lobbied the World Food Program for blanket-feeding in vulnerable areas in early 2006. (ECB, 2007, p 23)

This kind of institutional follow-up is a challenge for one-off exercises, in particular system-wide evaluations. But there is a tested model from the multi-agency Rwanda evaluation. A formal follow-up group – the Joint Evaluation Follow-up, Monitoring and Facilitation Network (JEFF) – was established, to continue disseminating the findings and recommendations and to monitor follow-up (JEFF, 1997). Ten years later, Danida took the unusual but important step of reviewing the longer-term impact and influence of the Rwanda evaluation (Borton and Eriksson, 2004).

Some interview respondents emphasised the importance of straightforward dissemination to target audiences, especially where learning is the overall purpose of an evaluation. This is not just about disseminating the report, but generating discussion. As an OCHA respondent said in relation to RTEs: 'workshops are more important than reading the report'. In the two most recent IASC RTEs in 2007, the report was written in country and discussed in a workshop before the evaluation team left. This creates a demanding schedule for the RTE team, but such immediate feedback and discussion has many benefits if the team can manage it. Involving senior managers in dissemination fora was also emphasised, especially in the UN system. For example, after the IASC RTE in Pakistan in 2007, the evaluation team leader met with the UN senior management team in New York, and 'had them focused for an hour'.

When we talk about use of an evaluation there is a tendency to focus just on the direct use of findings and recommendations, thus failing to recognise the indirect uses and benefits.[42] In joint evaluations where the process of bringing together different actors is a defining feature, the indirect uses can be very important. ECHO and WHO staff members emphasise and value the improved understanding and trust between them as a result of the Health Action in Crisis RTE initiative. Some of the ECB joint evaluations have a legacy of encouraging agencies in the field to work together more closely, with knock-on benefits. And one of the indirect benefits of the TEC has been to build the social capital of agencies involved, in terms of new skills and new partnerships (Houghton, 2006).

3.7 Conclusions and the future agenda for joint EHA

3.7.1 Conclusions

The rich array of different experiments in joint evaluations of humanitarian action is encouraging. This is no longer the domain solely of donor governments, the early champions of joint evaluations. UN agencies and some NGOs are now fully engaged. But it is still early days, and some efforts to promote and institutionalise a joint-evaluation approach have come and gone, for example the IHE initiative, despite evidence that joint evaluations help to build trust and social capital among participating organisations (our Hypothesis 1 – see Box 3.1). Reflecting the set-up of international humanitarian agencies, joint evaluations have so far been Northern and headquarters-driven (Hypothesis 2). The challenging next step is to make real progress in fully involving national stakeholders who have been poorly represented – this includes national NGOs and other organisations, and governments (Hypothesis 3). Involving the latter will be easier in natural disasters than in conflict-related humanitarian crises, especially if a government is an active party in the conflict. But there may be important learning for the humanitarian sector from joint evaluations in development, from work done by DAC to strengthen developing-country participation and from the Paris Declaration on Aid Effectiveness.[43]

One of our objectives in this meta-evaluation is to review the quality of joint evaluations of humanitarian action, comparing this with the quality of evaluations in the sector more generally. Our findings provide conclusive evidence that joint evaluations are overall of higher quality than single-agency evaluations (Hypothesis 8):

- their terms of reference are generally clearer and more usable
- consultation with local populations and beneficiaries is stronger (Hypothesis 4)
- more attention is paid to international standards (Hypothesis 6)
- the EHA criteria are more rigorously used.

Our working hypothesis that joint evaluations have more rigorous methodologies than single-agency evaluations (Hypothesis 5) is proven, but not across the board.

There are striking gaps and weaknesses in the joint evaluations reviewed, especially their attention to crosscutting issues such as gender equality, protection and advocacy (Hypothesis 7).

Our hypotheses that joint evaluations are more likely to address policy issues and locate their findings in the context of wider debates within the sector (Hypotheses 9 and 10) met with a mixed response. There is some evidence of this, for example in relation to UN-led humanitarian reform processes. But there were also missed opportunities which implied that a number of joint evaluations in our sample had not fulfilled this potential, despite the generally high quality of the evaluation reports.

The debate about whether joint evaluations replace or reduce the need for single-agency evaluations is an active but distracting one. In the words of one of our respondents, 'they are very different animals', and have different purposes. Joint evaluations can fulfil accountability purposes – and their better use of EHA/DAC criteria, revealed in this meta-evaluation, supports this role – but this may be at a different level from the accountability needs of a single agency. Accountability to peers, and to some extent to beneficiaries, through stronger consultation, featured in a number of the joint evaluations in our set. But if individual agencies need to be accountable to their funders in any detail, a joint evaluation may not fulfil this need.

Joint evaluations clearly complement single-agency evaluations by placing the response in the wider context, exploring how agencies work together, and addressing wider policy issues. One of the few occasions when a joint evaluation might reduce the need for single-agency evaluations is when a group of like-minded agencies come together to evaluate their work in a particular area, as the ECB and some UN agencies have done. In these examples, reducing the number of evaluation teams on the ground asking very similar questions of local communities, government officers and others is clearly a good thing. Fewer but more considered joint evaluations of this type might facilitate follow-up by reducing the overload on the humanitarian sector (see Sandison, 2006, p 141).

The value of joint evaluations for peer learning comes through strongly, from what has been written about them but also from agency interviews. If more joint evaluations had a deliberate utilisation focus, then their contribution here could be even greater. However, when deciding if it is appropriate to launch a joint evaluation, it is important to remember that evaluation is just one of a number of different

vehicles for promoting learning. The questions to be asked in making this decision, in order, include:

1. What is the scope and purpose of the exercise?
2. What are the key questions to be addressed?
3. What is the most appropriate methodology? Is a joint/collaborative approach appropriate?

Our sample for this meta-evaluation is dominated by joint evaluations of natural disasters, which are usually easier to negotiate because they are less 'political'. Learning from these experiences can play a useful role in preparing for future joint evaluations in more complex and sensitive conflict-related humanitarian crises. However, it may also be time to consider joint evaluations in some thematic and policy areas in the humanitarian sector that are relatively new and/or challenging to the international system (thus moving into some of the boxes in our typology table that are still blank). Possible candidates might include protection as part of humanitarian action, or livelihood support in the midst of a humanitarian crisis.

3.7.2 Learning points for use in future joint evaluations

The analysis in this meta-evaluation generates specific points of learning for application in joint evaluations of humanitarian action in the future.

The time it takes to do a joint evaluation is constantly underestimated. Joint evaluations do take considerably more time than single-agency evaluations:

- in setting up and negotiation, especially for a utilisation-focused evaluation
- for actually doing the work, especially the field work, as the scope of a joint evaluation is, by definition, greater than that of a single-agency evaluation.

Early pressures to get a quick result often prove to be false pressures, especially if the joint evaluation is intended to be policy-focused.

In-country buy-in is essential to the success of a joint evaluation, and should be central to a utilisation focus. Positive examples are emerging of how to do this

well, including pre-visits to set up the evaluation, early involvement of in-country decision-makers and in-country steering committees.

Negotiating the terms of reference can be time-consuming, but is worthwhile in terms of achieving good-quality ToR. Areas that still require strengthening are clearly identifying users, and being clear about the purpose of the evaluation (eg the intended balance between learning and accountability).

The pool of sufficiently skilled evaluators for joint evaluations is small compared with demand; this shortage could get worse if more joint evaluations are launched, implying a need to invest in evaluator capacity, especially in affected countries. For policy-focused evaluations, experience indicates the benefits of hiring policy analysts as team leaders or as team members.

Crosscutting issues of gender, protection and advocacy are poorly covered in joint evaluations. Attention given to protection and advocacy in humanitarian action is relatively new, but it is perplexing that gender is still poorly covered. There may be a need to provide some evaluation guidance in each of these areas.

A joint evaluation that is part of a wider institutional framework or relationship tends to benefit from better-established mechanisms for discussion and follow-up of recommendations. The implications of this are worth considering at the outset when agencies are considering joint evaluations in the future.

3.7.3 A future agenda

This meta-evaluation is the first stocktaking of joint evaluations in the humanitarian sector. The trend towards joint evaluations is expected to gather pace. ALNAP has been at the centre of much of the debate around joint evaluations, not least since it hosted the TEC Secretariat. It is the natural forum for continued exchange of experience around joint evaluations and for monitoring (and perhaps even coordinating) continued joint-evaluation activity. In short, ALNAP is well placed to provide ongoing leadership on joint evaluations in the humanitarian sector. The practical effects of this could include: ongoing facilitation of inter-agency learning around joint evaluations; guidance on joint evaluations of humanitarian action; dedicating a section of the ALNAP website to joint evaluations; tracking joint-

evaluation activity; and disaggregating the ERD into single-agency and joint evaluations[44].

Launching the TEC as the second-ever system-wide evaluation was timely, but also brave in the context of an increasingly complex international humanitarian system. Learning from the experience is considerable and should be written up to contribute to the growing body of knowledge on joint evaluations.

We recommend that a third system-wide humanitarian evaluation be considered in the next 18 months, this time focused on a significant but relatively 'forgotten'/ under-evaluated humanitarian crisis, for example in eastern DRC. This geographical focus would help to address international commitments to provide assistance according to need. This is an area in which the sector has much to learn; yet there would be less pressure to act fast, at the expense of process, compared with other high-profile humanitarian crises. It would provide an opportunity to apply the learning from the TEC immediately, while it is still fresh, for example on how to include local and national stakeholders, the early implications of being utilisation-focused, and taking steps to ensure greater policy focus. And it is a step towards system-wide humanitarian evaluations becoming a more regular feature of the landscape. We recommend that this proposal be discussed by the ALNAP membership. Along similar lines, there is potential value in a JE on a global humanitarian policy issue such as protection or livelihoods in conflict. This should also be considered by the ALNAP membership.

Joint evaluations are generally better at beneficiary consultation than other evaluations. They offer a valuable opportunity for exploring different and creative ways of consulting beneficiaries, comparing different approaches and capturing good practice. This should be the subject of an action-research project.

Many different management structures for joint evaluations in the humanitarian sector have now been tried and tested; this meta-evaluation has started to explore some of them. This challenging aspect of running a joint evaluation deserves more attention in future. A particularly important element of this relates to the structures and follow-up processes to ensure utilisation of the recommendations. We recommend a project that describes and analyses the pros and cons of different management structures, in relation to the typology of joint evaluations presented, to guide decision-makers in future in their choice of management structure.

Finally, our work suggests that JEs should be explicitly considered in agency evaluation strategies and policies, supported by an articulation of their relative strengths and constraints in relation to single agency evaluations. This can contribute to building awareness and understanding of the value of JEs within agencies, and to the institutionalisation of JEs.

Notes

1 This trend was noted in the 2005 meta-evaluation, which included 6 joint-evaluation reports in its overall sample of 30 reports.

2 The first system-wide evaluation had been carried out a decade earlier, evaluating the international response to the Rwanda crisis in 1996. The TEC was established in February 2005 as a multi-agency learning and accountability initiative in the humanitarian sector. It was managed by a Core Management Group of 14 organisations, and TEC staff members were hosted by the ALNAP Secretariat.

3 The sample is smaller than for previous meta-evaluations, reflecting a shift in emphasis in methodology to include a review of the literature, and also because most joint-evaluation reports are longer than other evaluation reports.

4 The two meta-evaluators are the authors of this chapter. Each has over 15 years of evaluation experience. One of the meta-evaluators also participated in the first four ALNAP meta-evaluations, which has ensured consistency of analysis and interpretation. Annexe 3.7 provides more background on the meta-evaluators.

5 Around 12 per cent of the evaluation reports lodged on ALNAP's ERD for 2005 to 2007 are for joint evaluations.

6 Darcy, J et al (2007) *External Evaluation of the Protection Capacity Standby Project* (mimeo).

7 See, for example, Dabelstein (1996) and Borton (2001).

8 The role of independently commissioned reports in the humanitarian sector should be noted, although they are not strictly evaluations – for example, the Kosovo Report produced by the Independent International Commission on Kosovo. These, too, can engage with the bigger picture and place the humanitarian response in context.

9 CARE International, Oxfam GB, World Vision, IRC, Catholic Relief Services, Mercy Corps and Save the Children are all members of the Interagency Working Group, an informal grouping of large INGOs. The ECB is a project created by that group, funded by the Gates Foundation, which has provided space for such joint exercises to take place.

10 Information on the reluctance to engage in system-wide evaluations based on personal communication with John Borton, ALNAP Coordinator at the time.

11 See for example, Samoff (2005).

12 See, for example, DAC (2006), ECB (2007) and IHE (2007).

13 For example, DAC (2005a), DAC (2006), ECB (2007), Buchanan-Smith (2007).

14 By transaction costs, we mean the costs of engaging with a consortium of agencies, such as increased staff time and travel costs.

15 'How actors work together', represented by the rows in Table 2.2, could continue to be developed as new forms of working together evolve in the humanitarian sector, for example around joint funding arrangements and pooled funds.

16 For example, the DFID-UNICEF evaluation of a programme of cooperation to strengthen UNICEF programming in humanitarian response in 2005.

17 Instead, the Tsunami Recovery Impact Assessment and Monitoring System (TRIAMS) was set up, but did not report on impact until after the TEC had been completed (http://www.ifrc.org/docs/pubs/updates/triams-presentation.pdf).

18 See DAC (2005a), which cites feedback from stakeholders interviewed for the

study. This also comes through in SIDA's evaluation manual.

19 Changes to the Pro Forma over the period of its use made comparison to meta-evaluations other than that of 2004 in this area difficult. However, results for the 2001–2003 meta-evaluations are similar to those for 2004. Percentage results are calculated on the incidence of ratings; where there was a disparate rating (eg one meta-evaluator rated good and the other satisfactory), the rating was divided between these two categories.

20 The 2005 meta-evaluation rated 60 per cent of evaluations as unsatisfactory or poor in this area.

21 Notes from an 'After Action Review' of the Tsunami Evaluation Coalition: Core Management Group (CMG) meeting, Geneva, 9 September 2005 (http://www.tsunami-evaluation.org/NR/rdonlyres/887A28A0-52C5-44AF-8376-01FB0260EFCE/0/tec_cmg_aar_050909.pdf).

22 DAC (2005b) sets out a preferred division of responsibility in joint evaluations.

23 Presentation at ALNAP December 2006 Biannual (http://www.alnap.org/meetings/dec06/Jointevaluationsworkshop2.pdf).

24 http://www.tsunami-evaluation.org/NR/rdonlyres/9DBB5423-E2EF-43AB-B6D2-2F5237342949/0/tec_lessonslearned_ver2_march06_final.pdf

25 A composite index of evaluation quality was developed using the following Pro Forma areas: DAC criteria, participation of beneficiaries, gender, protection and advocacy. The percentage of evaluations by mixed and international teams that achieved a good/satisfactory rating on this composite index was calculated. The former achieved 68 per cent, and the latter 58 per cent.

26 http://www.tsunami-evaluation.org/NR/rdonlyres/BC705D46-41B3-4521-B856-1D6D2EC2A38C/0/TECSurvey.pdf

27 Our own review of joint evaluations also falls into a similar pattern of focusing on international agencies, with no government interviews carried out or reports reviewed.

28 It is also interesting to note that two anthropologists familiar with the area were included in the team for Study 3 of the multi-donor Rwanda evaluation.

29 Comparison is made only between 2004 and 2006/07 because of changes in the Pro Forma in this area in 2004.

30 For calculation methods, see note 25 above.

31 The criterion 'coordination', which was introduced for this year's meta-evaluation, has been removed from these calculations.

32 ALNAP produced a guide to the DAC criteria in 2006, which was used by some joint-evaluation managers, so this may have contributed to improved results for joint evaluations (see http://www.alnap.org/publications/eha_dac/index.htm).

33 The interpretation of the OECD-DAC criteria might be different in system-/sector-wide evaluations, as they were originally designed for project level. For example, there is a difference between evaluating efficiency at the project level and in relation to a system-wide process of policy formulation.

34 One peer reviewer noted that part of the reason for higher quality of joint evaluations might be that there is an experimental effect at play – that is, the current spotlight on joint evaluations might have led to agencies being more diligent in their approach and practice regarding joint evaluations.

35 Notes from an 'After Action Review' held at the Tsunami Evaluation Coalition

(TEC) Core Management Group (CMG) meeting, Copenhagen, 21 September 2006 (www.tsunami-evaluation.org/NR/rdonlyres/D22B295E-0C88-46D4-A1027DECEBB1B935/0/AAR0609.pdf).

36 This section should be read in conjunction with Section 3.4.5 on gender equality.

37 See, for example, Buchanan-Smith (2003) and Borton and Eriksson (2004).

38 See, for example, Buchanan-Smith (2007).

39 However, this also represents a challenge facing the humanitarian sector, as responsibility, and therefore accountability, are not always clear or well-defined between humanitarian agencies. See Raynard (2000).

40 An example of this type of response to the high-profile multi-agency Rwanda evaluation and how it was handled has been well documented (see Borton, 2001; Dabelstein, 1996).

41 Some of this learning was recorded in the last meta-evaluation – see Wiles (2005).

42 See Williams et al (2002) on use of evaluation in the European Commission; and Sandison (2006).

43 See DAC (2005a) and *Evaluating the Paris Declaration Factsheet* (www.oecd.org/dac/evaluationnetwork).

44 This kind of role for ALNAP was encouraged at the 20th ALNAP Biannual meeting in Rome in December 2006, in a workshop dedicated to joint evaluations.

References

ALNAP (2001) *Review of Humanitarian Action: Learning from Evaluation*. London: ALNAP/ODI.

Beck, T (2006) *Evaluating Humanitarian Action Using the OECD-DAC Criteria* (ALNAP Guide). London: ALNAP.

Borton, J (2001) 'Doing Study 3 of the Joint Evaluation of Emergency Assistance to Rwanda: The Team Leader's Perspective' in A Wood, R Apthorpe and J Borton (eds) *Evaluating International Humanitarian Action: Reflections from Practitioners*. London, New York: Zed Books.

Borton, J and J Eriksson (2004) *Lessons from Rwanda – Lessons for Today. Assessment of the Impact and Influence of the Joint Evaluation of Emergency Assistance to Rwanda*. Denmark: Ministry of Foreign Affairs.

Borton, J et al (1996) *The International Response to Conflict and Genocide: Lessons from the Rwanda Experience. Joint Evaluation of Emergency Assistance to Rwanda. Study 3. Humanitarian Aid and Effects*. Copenhagen: Steering Committee of the Joint Evaluation of Emergency Assistance to Rwanda.

Broughton, B and S Maguire (2006) *Inter-agency Real-Time Evaluation of the Humanitarian Response to the Darfur Crisis.* A real-time evaluation commissioned by the United Nations Emergency Relief Coordinator & Under-Secretary-General for Humanitarian Affairs, Office for the Coordination of Humanitarian Affairs (OCHA).

Buchanan-Smith, M (2003) *How the Sphere Project Came into Being: A Case Study of Policy-Making in the Humanitarian Aid Sector and the Relative Influence of Research*. ODI Working Paper 215, London: ODI.

Buchanan-Smith, M (2007) *Joint Evaluations*. Report on a workshop on joint evaluations at ALNAP Biannual, Rome, December 2006. http://www.alnap.org/meetings/dec06/Jointevaluationsworkshop2.pdf

Dabelstein, N (1996) 'Evaluating the International Humanitarian System: Rationale, Process and Management of the Joint Evaluation of the International Response to the Rwanda Genocide' in *Disasters Journal* Vol 20, No 4, pp 286–294.

DAC (2005a) *Joint Evaluations: Recent Experiences, Lessons Learned and Options for the Future*. DAC Evaluation Network Working Paper. Paris: DAC.

DAC (2005b) *Workshop on Joint Evaluations. Challenging the Conventional Wisdom – the View from Developing Country Partners*. Nairobi, 20–21 April 2005, workshop report. Paris: DAC.

DAC (2006) *Guidance for Managing Joint Evaluations*. London: DAC.

ECB (2007) *What We Know About Joint Evaluations of Humanitarian Action. Learning from NGO Experiences*. Draft (v4), July, ECB mimeo.

EU (2007) *The European Consensus on Humanitarian Aid*. Joint Statement by the Council and the Representatives of the Governments of the Member States meeting within the Council, the European Parliament and the European Commission, December.

Houghton, R (2006) *TEC Lessons Learned*. Presentation to ALNAP Biannual, Rome, December 2006, workshop on joint evaluations.

IHE (2007) *Guidelines for Implementing Interagency Health and Nutrition Evaluations in Humanitarian Crises*. Version 1.0, August, mimeo.

JEFF (1997) *The Joint Evaluation of Emergency Assistance to Rwanda: A Review of Follow-up and Impact Fifteen Months after Publication*. London and Copenhagen: ODI and Danida

Lipsey, M (2000) 'Meta-analysis and the Learning Curve in Evaluation Practice' in *American Journal of Evaluation* Vol 21, No 2, pp 207–213.

Raynard, P (2000) *Mapping Accountability in Humanitarian Assistance*, Report presented to ALNAP at the biannual meeting in April 2000. London: ODI.

Samoff, J (2005) *Imaginative Observations and Muddling Through in Joint Evaluations. Observations drawn from Local Solutions to Global Challenges: Towards Effective Partnership in Basic Education*. Paper presented to 2005 Joint Conference of the Canadian Evaluation Society and the American Evaluation Association, Crossing Boarders, Crossing Boundaries, Toronto, 26–29 October 2005.

Sandison, P (2006) 'The Utilisation of Evaluations' in *ALNAP Review of Humanitarian Action. Evaluation Utilisation*. London: ALNAP/ODI.

TEC (2006) *Lessons on Multi-agency Evaluation from the TEC Process – March 2006*. London: ALNAP.

TEC (2007) *Summary of Responses to the TEC Survey Questionnaire*. TEC website. http://www.tsunami-evaluation.org/NR/rdonlyres/BC705D46-41B3-4521-B856-1D6D2EC2A38C/0/TECSurvey.pdf

UNEG (United Nations Evaluation Group) (2005) *Standards for Evaluation in the UN System*. New York: UNEG.

WFP (2007) *Summary Report of the Mid-Term Evaluation of the India Country Programme (2003–2007)*. Rome: WFP (WFPEB.1/2007/7-B).

Wiles, P (2005), 'Meta-evaluation' in *ALNAP Review of Humanitarian Action in 2004*, London: ODI.