CHAPTER 4
OVERVIEW OF THE ANALYSIS AND NEXT STEPS

4.1 Introduction

This chapter combines the analysis of the principal findings and meta-evaluation contained in Chapter 2 and Chapter 3, thus covering forty-nine individual evaluation reports, five synthesis reports and one summary report.

In synthesising the principal findings from both the Kosovo and non-Kosovo sets, section 4.2 highlights the differences, considers likely contributing factors, and the implications of the overall findings for the humanitarian system and for ALNAP's own agenda.

As well as combining the meta-evaluation from the two preceding chapters, section 4.3 adds to the analysis of report quality through an application of the preliminary proforma to the Kosovo set. It concludes with a consideration of the next steps needed to strengthen the role of evaluation within the humanitarian system and, again, the implications for ALNAP.

4.1.1 Contextual differences between the sets

The two sets of reports differ significantly in terms of the contexts in which the operations being evaluated took place. Of the complex emergency/natural disaster contexts covered by the non-Kosovo set only East Timor displays similarities to Kosovo, in respect of the efforts made towards peace enforcement, the setting up of an interim administration, nation-building, etc. Only one East Timor evaluation was available for inclusion in this Annual Review.

Detailed analysis of per capita expenditures in each of the operations evaluated was not readily available to the authors. However, perhaps the most striking difference between the Kosovo and non-Kosovo sets is the sheer scale of resources deployed in Kosovo, far outstripping those deployed in the other contexts (see 4.2.1 below).

Despite such obvious differences between the contexts, many of the Kosovo evaluations are comparable to the non-Kosovo evaluations in that they generally shared the same criteria, were generally undertaken by consultants, and focused on one particular area or sector.

4.2 Principal Findings & Issues Raised by the Reports

4.2.1 Directions of funding

The pronounced bias towards Kosovo in the allocation of funds by the humanitarian system has been raised by other publications (e.g., IASC, 2000), and discussed by one report as follows:

'While donor governments gave US$207 per person through the 1999 UN appeal for Kosovo, those in Sierra Leone received US$16, and those in the Democratic Republic of Congo little over US$8 ... The cost of Camp Hope in Albania for 3,000 refugees for two months was equivalent to the 1999 UN appeal for Angola ... In Africa for the past 20 years, refugees have often had to survive on lower than basic food requirements. Kosovo refugees in some cases were receiving Turkish delight in

complimentary food packages. In the emergency, 80,000 refugees were flown to other countries in Europe, something that would never be conceived of in an African crisis,' (DEC, 2000c vol 1: p77).

Such disparities clearly show that the international humanitarian system is strongly influenced by factors such as national interest, geopolitical strategising and media focus. As a system it falls significantly short of meeting the criteria necessary for the use of the term 'humanitarian', such as impartiality, independence and considerations of proportionality, as contained in International Humanitarian Law and the Red Cross/NGO code of conduct.

A necessary first step in equipping the system to counter such biases, is the monitoring of humanitarian expenditure on a systematic basis. The Inter-Agency Standing Committee's publication *Global Humanitarian Assistance 2000* (IASC, 2000) represents a welcome development in this regard. It was obliged, however, to make use of datasets that were not always comparable or consistent, and it is not clear whether it will be produced on a regular basis. Efforts need to be made by the DAC, OCHA and the system generally to strengthen the datasets essential to such monitoring.

A second step would be work to develop measures of humanitarian need to enable agencies involved in funding and coordinating a response to determine at an earlier stage than at present whether the humanitarian needs of affected populations (however defined) were being met, not met or significantly exceeded. Initial work in this field is planned by the Overseas Development Institute and Tufts University. Although it will inevitably encounter conceptual and technical difficulties, it can build on accepted existing approaches in the nutrition, food security, and public health sectors, and make full use of the widely recognised Sphere Minimum Standards in Disaster Response (Sphere, 2000).

Another funding-related issue emerging from the reports is that a significant percentage of the funding being evaluated, perhaps between 30 and 50 per cent, concerned rehabilitation activities rather than activities which addressed 'acute human suffering'. These proportions are based on an impressionistic assessment of the activities evaluated and are very approximate. Nevertheless, they raise important questions as to the overall appropriateness of impact indicators such as 'number of lives saved' for an assessment of the overall intervention. In addition they point to the need for datasets on humanitarian expenditure to make a clearer differentiation between relief and rehabilitation.

4.2.2 The results of humanitarian action

Judged on their own terms, interventions were assessed by the evaluations as being largely successful at meeting requirements and their stated objectives, with inputs leading to appropriate outputs. This in itself is a remarkable achievement. The efforts of agency personnel, which appear to be one of the most important factors contributing to success, need to be highly commended, given the context and their difficult working conditions. At the same time, the claims of success need to be qualified.

As Chapter 2 and Chapter 3 point out, the stated objectives of interventions were often vague. The measures by which these objectives were assessed – mouths fed and lives saved – could have been more focused and disaggregated. In particular the Kosovo set tends to focus more on organisational issues and coordination or, at a lower level, on the results chain, either inputs or outputs. We learn for example that funds were used to rebuild housing, but not the uses of that housing (Tearfund, 2000).

Evaluation of impact – that is the real difference the activity has made (for a full definition see OECD-DAC, 1999) – was required in many terms of reference, but commissioning agencies appear to have been satisfied with reporting on effectiveness or input instead. The challenge in this case is promoting

an understanding of the difference between effectiveness and impact, and ensuring that both areas are adequately addressed in evaluations.

Interventions were also assessed by the evaluations as being for the most part relevant to the affected populations in both the Kosovo and non-Kosovo set. In this case the standard of relevance does not appear to include systematic assessment of the views of the affected population, undermining overall findings.

In addition, the Kosovo and non-Kosovo evaluation reports were consistently weak in key methodological areas (notably triangulation, consultation with the affected population, and lack of attention to gender). They were for the most part opaque in methodology, to a much greater extent than is the norm in evaluations of development cooperation or the general evaluation field.

Few of the reports raised the issue of attribution adequately and only one attempted counter-factual analysis. The Kosovo reports in particular were unable to isolate the main causes of results – e.g., what was due to the good health and coping strategies of refugees, and the support provided by host families, and what to humanitarian action. Determining the factors responsible for a particular result is a major function of evaluation, and so this inability to isolate the main factors represents a significant shortcoming. This failure to establish credible indicators of success, combined with weak methodological approaches, must call into question the validity of the conclusions reached by many of the reports.

4.2.3 Coverage

The majority of evaluation reports concluded that coverage of affected populations and vulnerable groups within the programmes evaluated was at the least adequate, and that resources were not in general diverted. That resources were provided to vulnerable groups is a significant achievement and should help counter criticism in this area. On the other hand, both the Kosovo and non-Kosovo sets illustrate that the issue of coverage needs to be accorded more careful consideration by evaluations. In the Kosovo case a number of reports point out that the major focus of interventions in Albania and Macedonia was the refugee camp populations, to the exclusion of attention to refugees in host families. In the case of the non-Kosovo reports, the lack of attention to IDPs is similarly noted. Another issue relating to coverage is that while assistance may for the most part be reaching vulnerable groups, this does not mean that all of the vulnerable population is being reached, a point not adequately considered in a majority of the reports.

4.2.4 Lack of attention to protection

A common theme from the evaluation reports is that humanitarian agencies are not giving adequate attention to protection, humanitarian space and human rights in complex emergencies. This point comes out most clearly in the seven DANIDA evaluations which note to differing degrees the importance of protection as a key strategy. It is common for agencies to focus on the provision of material assistance and services such as health, food and shelter, leaving protection to agencies such as UNHCR and ICRC that have specific protection mandates. While UNHCR (2000a) notes the creditable performance of that agency in providing protection to refugees in the countries of asylum around Kosovo, Chapter 3 argues that lack of protection, and particularly armed protection of vulnerable minorities by the humanitarian system and the international community within Kosovo, should have been a main focus of those evaluations given the human-rights violations there.

Within the humanitarian system there is now greater recognition than before that all humanitarian agencies have a role to play in relation to protection. A range of activities is underway which should contribute to an improvement in the situation. For instance the ICRC Protection Workshop Process, begun in 1996, has held four workshops so far (e.g., Von Flüe and de Maio, 1999), and IASC will shortly publish a practitioner's field practice manual (Paul, 2001).

However, debate continues as to what constitutes an appropriate role for different types of humanitarian agency. Paul (1999) notes that in considering their role in relation to protection, NGOs are often concerned with how their involvement in protection activities might affect their ability to carry out their mission; jeopardise their principle of neutrality; and risk their access to areas or countries controlled by groups/authorities antagonised by their protection activities. Addressing the issues raised by the evaluations is likely to require genuinely collective system-wide efforts to establish common understandings and approaches.

In addition, there is a parallel need to develop standards and criteria for use in the evaluation of protection. For instance UNHCR (2000a) notes that UNHCR was criticised for being both too timid and too forceful in its relations with the Government of the Former Yugoslav Republic of Macedonia and in respect of the initial refusal of asylum to refugees. Evaluations can support the development of such standards, and the UNHCR evaluation (2000a: Chapter 6) is itself an example of good practice in the use of standards for assessment of protection (even if limited to refugees from Kosovo and not human rights violations in Kosovo).[1]

4.2.5 Coordination

Coordination has a well-established reputation for being amongst the most problematic performance areas within the humanitarian system. That both the Kosovo and non-Kosovo evaluations support this view comes as no great surprise. Kosovo provides one of the most telling examples, where UNOCHA was not called on to play a main coordination role and UNHCR, designated as the UN Lead Agency, was to an extent marginalised by NATO and bilateral donors for what were largely political reasons (see Chapter 3 and UNHCR 2000a). There are also many other examples in the non-Kosovo evaluations (see Chapter 2).

Debates on the most appropriate modalities for coordination in the system continue, with little advance on those that began in the 1970s. While some praise and support OCHA, others criticise and undermine it. Likewise, some argue strongly in favour of the Lead Agency model, while others argue equally strongly against it. The adage that 'all agencies want coordination but none want to be coordinated' remains useful in explaining the coordination problem. However, the extraordinary persistence of the problem can probably only be explained in terms of structural factors such as the large number of separately funded, separately governed agencies typically involved in any single operation. Short of radical measures to reduce the number of agencies or channel funds through a limited number of central coordinating bodies, the system's structure will probably only achieve significantly improved coordination if it is able to develop an appropriate framework of incentives and disincentives for agencies to coordinate. Currently, the rewards for good coordination and the penalities for not coordinating are just too weak to have a noticeable effect.

The persistence of the problem also raises issues for evaluation. Are evaluations dealing adequately with coordination as a central problem in the system? Almost certainly not, as their primary focus is on the performance of single agencies with coordination frequently regarded as a secondary issue. This argues for more collaborative evaluations. Are the necessary coordination standards and

indicators in place for use by evaluators? Again the answer is negative, and this lack needs to be urgently addressed.

However, the picture provided by the evaluations on coordination is not all gloom. Examples of good practice are revealed in, for instance: OCHA's Mine Action programme in Afghanistan (DANIDA, 1999b); actors in the 1998 Bangladesh floods (DEC, 2000a); and the EMG in Albania (DFID, 2000). A common factor in each is a significant host country input – from Afghani staff in the Mine Action programme, to Bangladeshi government and NGO input, to the location of the EMG in the Prime Minister's Office in Albania. This relates to the need for good partners on the ground, a point agreed by all the reports that covered the issue. Partners were usually conceptualised in terms of INGOs, NGOs or multilateral institutions, but can clearly also include different levels of national government.

4.2.6 Linking relief, rehabilitation and development

While the non-Kosovo set of evaluation reports raises LRRD as another of the most difficult areas, half of the Kosovo evaluations do not. Their focus is on the relatively brief period when the main refugee population was outside Kosovo. However, the seven ECHO Kosovo evaluations find that LRRD remains a major concern for ECHO, and the main points raised in the non-Kosovo reports are very similar to those in the Kosovo reports that do cover the issue:

- a lack of policy and strategy to guide LRRD;
- a lack of communication within and between agencies;
- an inability to plan adequate exit strategies.

These results are consistent with a major recent study on overall (rather than just humanitarian) assistance to post-conflict recovery:

'Despite ostensible good intentions, too often aid promised has not been committed, aid committed has not been delivered, and aid delivered has arrived too late. Moreover, the planning and implementation of reconstruction aid has frequently suffered from inadequate preparation, poor coordination and lack of perseverance,' (Forman and Patrick, 2000, p1).

The so-called 'grey-zone' between relief and development is not just a problem for the agencies being evaluated, it is also a problem for evaluation. Given that a significant proportion of humanitarian action budgets are being expended on both rehabilitation and development activities, evaluation reports need to be more specific in their recommendations concerning LRRD. An example of good practice in this regard is ECHO (2000p) that recommends a timeframe for hand-over; indication of the names of agencies intended to take over work; and a mid-term review of the hand-over process. A useful exercise would be the collection and dissemination of good practice in the evaluation of LRRD by ALNAP. Ideally this would be undertaken in collaboration with the Conflict Prevention and Post-Conflict Reconstruction Network.[2]

4.2.7 Coping strategies

There is evidence from the reports that the issue of coping strategies is on the agenda of some agencies, even if it is still dealt with in a sporadic fashion. While the Kosovo and non-Kosovo complex emergencies/natural disasters were very different events, affected populations used similar kinds of coping strategies. Primary among these strategies was mutual support. In the non-Kosovo cases Chapter 2, Box 2.3 has already referred to the importance of neighbourhood solidarity in response to Hurricane Mitch, and formal and informal association in East Timor.

Most Kosovo evaluation reports note the reliance on host families and remittances from abroad, and some (e.g., ICRC/IFRC, 2000) provide quite detailed information on feelings of mutual solidarity, as well as conflict, between host families and refugees. What the affected population was itself able to organise in the Kosovo case clearly outweighed the effects of humanitarian action, a point also made by the DEC with reference to natural disasters in Orissa and Central America (DEC, 2000, 2000b). In the Kosovo case, agencies failed to support adequately these coping strategies, partly because of their focus on refugee camps (see Chapter 3). A similar point can be made about the non-Kosovo interventions. A challenge for the future is to understand the different coping strategies of members of the affected population, and build on these during interventions.

4.2.8 Knowledge of context

The evaluation reports demonstrate a major gap in agencies' understanding of the context in which they were intervening. In Kosovo this was one of the most consistent messages from the reports, where the nature of 'ethnic' conflict in particular was inadequately considered. Without an adequate appreciation of context it is difficult, if not impossible, to design appropriate programmes. This is linked in turn to agencies' inadequate attention to coping strategies, where understanding how local and affected populations act in times of crisis is a major factor in understanding local context.

Why do agencies have such difficulty in gaining an understanding of the context in which they are operating? Among likely contributing factors is the lack of analytical tools for use by agencies, though the gap in relation to war economies is now being filled (Le Billon, 2000). Another factor is inadequate use of the capacities of existing local staff. However, the principal factor must be that agencies are simply not according sufficient priority to the development of the necessary analytical capacity, either within their own organisations or elsewhere in the system. Developing the capacity requires an investment that many agencies find difficult to justify or sustain when in direct competition with provision of assistance, or investments in what are perceived as more useful 'sharp end' activities such as logistics.

The reluctance of agencies working in challenging and dangerous environments to make such investments, may highlight the difficulties faced by a system composed of so many separate agencies. Some investments are seen to be uneconomical at individual agency level. The necessary economies of scale may be achieved by a number of agencies sharing costs. One example of such an initiative is the 'Learning Support Office' model (see Endnote 5,Chapter 3) being developed by ALNAP to provide services to a range of agencies – UN, Red Cross, NGO or others.

4.2.9 Lack of preparedness and rapid response capacity

Another consistent message from the reports is a lack of preparedness and capacity to respond expeditiously to sudden-onset humanitarian needs. In terms of preparedness, this was not just an issue of a lack of information but also of a failure to act on available information. Preparedness does not appear to be an area bilateral donor organisations are overly willing to fund, preferring instead *ex post* interventions. This is perhaps due to the low profile nature of preparedness work and the possibility that funding preparedness may not have results.

Following their experience in Kosovo and elsewhere, several agencies are in the process of setting up rapid response capacities with dedicated personnel or rosters of personnel deployable at short notice, linked in some cases to stockpiled materials. According to the ECHO evaluation (ECHO, 2000o) UNHCR is currently undertaking a programme which includes the following:

- enhanced systems of personnel recruitment and emergency deployment with 30 experienced personnel members, drawn from HQ and the field, on standby for emergency deployment within 24 to 72 hours;
- set up of an emergency pool of senior managers, to lead or coordinate large-scale and complex emergencies;
- intensive emergency training, especially at the senior and middle-management levels;
- external stand-by arrangements with a number of key partners strengthening emergency reserves;
- a database to determine the emergency capacity of NGOs.

However, agencies have made such investment in the past and, over time, the capacities have been eroded in periods of budget cuts, reorganisations or simply as a result of changed priorities. It remains to be seen whether the agencies will be able to sustain such investments this time around.

Once again the situation presents issues for evaluations. Standards as to what constituted an adequate level of preparedness in a given situation do not seem to have been available to evaluators, with the exception of fairly limited areas, for example, the need to have emergency teams on site within a certain period. The collection and dissemination of good practice case studies would be of assistance, and could be developed into guidance on how to evaluate preparedness.

4.2.10 Inadequate monitoring systems

The attention given by agencies to monitoring is found to be limited in almost all the evaluations. They also consistently lament the lack of systems that could provide data about results. Chapter 3 makes the point that an unmonitored programme is unevaluable in relation to standard performance measures.

Why monitoring is so weak is not clear, but may however relate to the difficulties of establishing effective systems in the first phase of response; the slow trickle down of performance measurement to humanitarian action; the idea that monitoring is not relevant in situations where life is in danger; and a lack of technical skills of agency staff.

In their monitoring recommendations, the reports often repeat evaluation mantras such as 'set clear objectives' and 'develop indicators', without recognising the constraints to such activities. DRC notes:

'With events moving so fast all the time in the Kosovo emergency it sometimes appeared pointless to those involved to use indicators of "success" and "failure" since the criteria for them kept changing. What at one stage appeared as a failure might later turn out as an achievement, or in any case as a learning experience,' (DRC, 1999: p23).

Evaluators and evaluation guidance consistently underestimate the constraints to, for example, setting clear objectives – e.g., the need to negotiate with a range of actors; political constraints; pressures to allocate and disburse funds; and the need for specific technical skills. Unless constraints are recognised by evaluators and commissioning agencies alike, project managers are unlikely to pay adequate attention to general recommendations relating to monitoring and evaluation systems.

4.2.11 Human resources and management issues

The picture that comes across from most of the reports is of under-trained, under-resourced staff doing their best in difficult conditions**.** Of all the problems noted, this is one of the most important to rectify. Good performance crucially depends on the 'right woman or man, in the right place, at the right time', to paraphrase DFID (2000). The numerous recommendations on human resources fall into three main categories:

- the need for more senior, better qualified personnel to be present at the beginning of the complex emergency/natural disaster;
- the need for personnel to be able to carry out adequate needs assessment including consultation with the affected population, gender and political analysis, and monitoring;
- less staff rotation.

One level of constraints relates to incentive structures within agencies. For example, a general lack of incentives to include gender, or to volunteer for field positions during complex emergencies (UNHCR, 2000a). A second level relates to lack of investment in personnel, including adequate training. In a culture of budget cuts and continuous organisational change, these constraints are difficult to overcome. This is, however, perhaps the area most likely to produce immediate improvements in humanitarian action. Future Annual Reviews might usefully pay attention to capacity development, and report on good practice and advances in this area. They might also monitor whether agencies are meeting their commitments, for example the ECHO Agenda for Action (ECHO, 2000).

People in Aid recently published the audited results of the two-year pilot to implement the People in Aid Code of Best Practice in the *Management and Support of Aid Personnel* (Davidson and Raynard, 2001). Seven of the original twelve agencies were recognised as fulfilling the Code's conditions. Their commitment to the Code will be re-audited in three years. The framework is, therefore, now established for many more agencies to commit themselves to the Code, and future Annual Reviews will report on the number of agencies making this commitment.[3]

4.2.12 Moral dilemmas in the Kosovo case

Perhaps the major difference between Kosovo and the other complex emergencies/natural disasters was the role of NATO in relief activities and the relationship of humanitarian actors to NATO. This brought greater attention, if not greater focus, to an ongoing dilemma as to the extent to which humanitarian action can be independent from political pressure (ECHO, 2000o, and see discussion in Chapter 3). Some Kosovo reports consider in detail the moral aspects of cooperation with a military partner (NATO) fighting a war unsanctioned by the UN Security Council (e.g., ECHO, 2000o; MSF, 1999c; DEC, 2000c). For the most part there is an acceptance that, given geopolitical realities, cooperation was inevitable (e.g., ECHO, 2000o). The focus of many of the reports is thus mainly on the mechanics of cooperation, rather than on ethical and moral issues. 'Saving lives' became the justification of cooperation with NATO: 'The humanitarian community generally accepted that the imperative of saving lives required cooperation with NATO,' (UNHCR, 2000a: p135). However, even the more sophisticated Kosovo reports, such as the one just quoted, do not appear to have addressed the issue of cooperation in sufficient depth and with sufficient balance.

The unsatisfactory nature of 'lives saved' as a non-disaggregated indicator is discussed throughout this *Review*. We need to know, as Chapter 3 points out, whose lives were saved by this cooperation and how? what were the effects of not providing protection to the population inside Kosovo? and, what were the unexpected or unplanned consequences?

A notable exception to the lack of attention to moral and ethical issues is the MSF evaluation that notes while discussing cooperation with NATO:

'MSF stood out as one of the most outspoken and critical NGOs, but was a lot of times the only one to express concern ... For many NGOs 'independence' is not on the agenda. For some it presents a concern but they are trapped on the necessity of institutional financing to be able to survive and are not willing to endanger their position by challenging decisions. Last not least (*sic*), there are NGOs which

are perfectly aware of the concept of independence, but give it no importance, what so ever. They have been created as an operational branch of their government,' (MSF, 1999c: p25).

The point for this Annual Review is not that there is a correct opinion about cooperation with NATO, but, that commissioning agencies should give clearer guidance on the need to pay greater attention to the moral and ethical issues likely to be encountered, when preparing the terms of reference for evaluations.

4.2.13 Shelter and housing

Housing is noted as one of the main problem areas of intervention in the non-Kosovo reports. In particular they highlight the lack of appropriate material and siting, poor targeting, and poor management and coordination. In contrast, there is surprisingly little critical analysis of the Kosovo housing-related programmes. Information as to the recipients of housing support in the Kosovo intervention, and the use made of that support, is negligible. Targeting issues and the relevance of the five-part categorisation used for allocating housing resources are discussed analytically in Tearfund (2000), but not elsewhere. Otherwise we learn in passing, for example, that many Kosovars own more than one home (ECHO, 2000o), but not the implications of this for the housing programme.

As a consistently problem area across a number of interventions, there is a clear need for a collation of lessons and good practice, as well as for the development of training modules/courses in this area.

4.2.14 Participation

About one-third of the non-Kosovo reports noted a good level of participation of the affected population, at least in the implementation of interventions, and about half of the evaluations considered participation. However, in the Kosovo evaluation reports, little is said about the level of participation. We learn in passing, for instance, that there was little participation of the affected population in the construction of camps by the military (UNHCR, 2000a) and extensive self-help and mutual support in the reconstruction of houses in Kosovo (ECHO, 2000o). Also that (DEC, 2000c: p34): 'It was not apparent, however, that a priority was given to beneficiary consultation ...'. Again the focus on political and organisational features in the Kosovo evaluations may have excluded attention to the role of the affected population.

Despite the introduction in 1994 of the Red Cross/NGO code of conduct committing signatory agencies to finding ways 'to involve programme beneficiaries in the management of relief aid', there remains wide variation in practice and continued criticism on the inadequate involvement of beneficiaries and affected populations. The bulk of humanitarian action is for the most part still a 'top-down' affair and 'downward accountability' to the beneficiaries and affected populations remains very weak. Recent progress by ALNAP in establishing a project to develop good practice and guidance on consultation with, and participation by, beneficiaries and the affected population in the planning, management, monitoring and evaluation of humanitarian programmes is to be welcomed (Endnote 1, Chapter 1).

4.2.15 Humanitarian evacuation and transfer programmes

The humanitarian evacuation and transfer programmes (HEP) in the Kosovo conflict were distinct new features. It is hard, however, to judge how relevant these programmes are to the development of policy and practice in the humanitarian system, for as the UNHCR evaluation report notes: 'The constellation of strategic and political interests that made evacuation programmes possible in this case is unlikely to

recur frequently,' (UNHCR, 2000a: p xii). The same evaluation notes that during the implementation of HEP, UNHCR's attempts to abide by international agreements and ensure refugees' rights to appropriate asylum in countries other than that of first asylum, were overruled by bilateral actors for political and visibility reasons.

4.3 The quality of Evaluation Reports

4.3.1 Overview

This section is based on the assessment of the 49 individual evaluations against the preliminary 'quality proforma' developed by ALNAP for this Annual Review (see Chapter 1 and Annex 1).

About 20 per cent of reports were considered to be using good practice in most areas covered by the proforma. In the remaining 80 per cent there were many elements of good practice, but on balance the reports were weak in about half the proforma areas.[4] Both Kosovo and non-Kosovo reports were weaker on the substantive areas of methodological rigour; attention to gender and international standards; and consultation with affected populations. This points to system-wide gaps that need to be addressed by commissioning agencies. Reports were stronger in the provision of contextual information (particularly the non-Kosovo evaluations); relevance and clarity of **terms** of reference; legibility; ease of access; and in crafting recommendations.

Almost all evaluations took a conventional evaluation approach, attempting to combine interviews with agency staff at HQ and in the field; with affected populations (in a minority of cases); document review; and observation of projects. There was little explanation or justification of methodological approaches taken. While there may have been attempts to cross-reference between the different information sources in a few cases, the outcomes were never explained adequately, for example, in terms of how conclusions were reached.

Reports assumed that their approach was 'objective', but did not provide discussion of potential bias. Biases were evident across the 49 reports – e.g., lack of attention to gender and affected population consultation. Although no evaluation or assessment, including the present one, can be free from evaluator bias, the point is that biases should be made clear and, in particular, how they have influenced the approach and the analysis of the evaluation.[5]

Another key question raised by the reports, and discussed throughout this Annual Review, is the appropriateness of impact indicators for humanitarian action. The answer seems to be, particularly in the case of complex emergencies, that indicators need to cover both 'traditional' areas such as the impact of shelter, food aid and support to health, as well as 'new' areas including protection, humanitarian space, human rights and advocacy. In addition, decisions will need to be made as to how to weight these indicators (to avoid the 'well fed but dead' syndrome noted in one DANIDA report) to determine overall impact. This is further discussed in the Next steps section below.

Using Borton and Macrae's (1997) synthesis as a very approximate baseline, there does appear to have been some progress made in evaluation quality over the last five years, including greater consistency in approach and coverage of key evaluation topics. The OECD-DAC (1999) Guidance, its earlier companion aimed more at field workers (Hallam, 1998) and papers produced by ALNAP itself (liberally referenced and used in the reports) have played a significant role in this improvement. Still, this progress has been from a low base, and much remains to be done (see 4.5 below).

The sections below follow the outline of the proforma. More detailed information on the non-Kosovo reports can be found in Chapter 2.

4.3.2 Purpose and focus of evaluation

The majority of the 49 reports have a mixed lesson-learning and accountability focus, usually roughly equal. This duality attempts to meet the needs of different audiences and terms of reference requirements. However, few of the reports consider the tensions, creative or otherwise, between the two (see Chapter 1).

4.3.3 Constraints

The Kosovo reports note similar constraints to the non-Kosovo set – e.g., lack of data and the complexity of the intervention in relation to the time available to evaluate it. The ECHO Kosovo evaluations and DEC (2000c) note that the refugee emergency was long over by the time the evaluation teams arrived, raising questions of recall and attribution. The reports did not tend to linger on constraints, usually noted in passing. These passing comments, and the few reports that did deal with this issue in more detail, highlight the fact that evaluations of interventions, in response to complex emergencies and natural disasters, face considerable constraints. Some are specific to humanitarian action, such as security issues and lack of access to areas of conflict or disaster. Others are more general in the evaluation and research fields: conflicts among the team or with the commissioning agency; inadequate direction and support from the commissioning agency; reliance on interpreters and agency personnel or the commissioning agency for logistics; and lack of time.

For purposes of transparency and credibility, reports would do well to expand their discussion of these constraints and illustrate their impact on the coverage and outcome of evaluations. For example, Kosovo evaluators must have been almost entirely dependent on interpreters in their discussions with the affected population, but none of the reports makes mention of this as a potential constraint.

4.3.4 Terms of reference, team composition, time allowed

Overall terms of reference were adequate where information provided on the nature and objectives of the intervention and clarity of purpose were concerned. Whether team composition was adequate is impossible to say, given the lack of details provided on this. Commissioning agencies tended to draw on a fairly narrow range of institutions and individuals, presumably known quantities. In total, 36 of the individual evaluations were carried out by expatriates, 12 by a mix of host-country citizens and expatriates, and 1 by host country citizens. Of the 55 evaluators involved in the Kosovo evaluations, 52 were expatriates, and 53 were independent consultants (with 2 being agency staff, one from UNICEF and one from WFP).[6] On the basis of the information provided in the reports, it is not possible to say if these consultants had any prior experience in the region, but it can be assumed in the majority of cases that they did not. The lack of contextual understanding and social analysis, for the Kosovo evaluations in particular (see Chapter 3), is a direct result of this hiring pattern by commissioning agencies. The parallel section in Chapter 2 expands on the discussion of team selection, recommending that a mixed team of expatriates and host-country evaluators should be used.

Allocation of time to undertake the evaluation effectively does not appear to have been adequate in the majority of cases. This appeared to be a particular constraint in relation to consultation with affected populations, undermining overall credibility. Commissioning agencies should ensure that sufficient time is allowed for consultation with the affected population, and ensure that this consultation takes place effectively. The evaluation team or team leader can be actively involved in planning this element of the terms of reference.

4.3.5 Information on context and intervention

While information on context and the intervention was adequate overall, the Kosovo and non-Kosovo reports display some differences. The Kosovo reports, with some notable exceptions (e.g., DEC, 2000; UNHCR, 2000), are weaker in the provision of context, and, as Chapter 3 notes, presented 'crises' without 'conflict', and 'ethnic' conflicts without adequate analysis of ethnicity. The difference may be accounted for by the fact that the non-Kosovo evaluators were more likely to be specialists in the country/region of intervention.

The reports did not tend to make good use of contextual information. They provided a narrative history of events and actors, but often failed to make necessary linkages between these events and the intervention to illustrate constraints and opportunities. Context was included, perhaps because it is conventional to introduce evaluation reports with it or because it was demanded by the terms of reference. Terms of reference should therefore make clearer the purpose of providing contextual information and the need to link it throughout the analysis. The conclusions of Chapter 2 and Chapter 3, and the assessment against the proforma, also demonstrate the need for evaluation standards in this area, as standards currently differ widely between agencies and individuals.[7]

4.3.6 Methodology and transparency

Few of the reports assessed provided adequate detail on, and justification for, the use of the methodology selected. Given that the choice, development and defence of methodology is a major component of evaluations, this is a significant shortcoming. This is not to say that the methodologies chosen necessarily lacked rigour, only that the reader is not given sufficient information to decide about the appropriateness and strength of methodology. Greater transparency in this area – for example, providing the numbers of affected population consulted (including their gender, socioeconomic grouping and ethnic background) and details of how they were consulted – will add to the evaluation's credibility (see 4.5 below).

Cases of evaluations being potentially biased because they relied mainly on one information source, were rare. Conventional approaches were generally used – e.g., interviews with agency/government staff; document review (although often not integrating the findings from secondary documents into the text); interviews with the affected population; and project visits. There was very little experimentation or use of techniques such as Participatory Rural Appraisal. As noted, while evaluations did use a range of sources, they rarely cross-referenced (triangulated) between them to increase reliability. Evaluation of humanitarian action has some way to go to catch up with techniques in these areas, already commonly used in the evaluation of development cooperation.

Given problems of attribution, it might have been expected that evaluations would make use of control groups, and consultation with the affected population not covered by the intervention, as a means of verifying impact. Only one of the 49 individual evaluations did. The Kosovo conflict would, in particular, appear to be a case where control groups were needed, even if this was made difficult by the reverse movement of refugees back to Kosovo. This is because the affected population did so much for themselves, which means that the impact of external interventions may have been quite limited. Unless we know how well non-recipients of humanitarian action fared, the results of this massive humanitarian programme cannot be plausibly ascertained.

4.3.7 Consultation with beneficiaries and affected population

Beneficiary perspectives are sorely lacking in both sets of reports. Comparatively, there is less heard from the affected population in Kosovo, where only two reports were assessed as being satisfactory or

better in this area. This may be because much of the affected population had already returned to Kosovo, while the evaluations focused mainly on operations outside Kosovo. Those evaluations that did focus on operations in Kosovo do not reveal better performance in terms of consultation with the affected population.

Only six evaluation teams, including those undertaking three of the DEC evaluations (DEC, 2000, 2000b, 2000c) consulted adequately with the affected population *and* included adequate methodological details about this consultation. Given the limited information provided in many evaluations, it is impossible to say who was actually consulted; how they were consulted; why they were selected; and what use was made of the information gathered.[8] These evaluations may have consulted adequately, but the reader is often left with statements such as: 'In the field the team spent as much time as possible with the beneficiaries … ' (ECHO, 2000), and little else. Even in those cases where there was a good level of consultation, the information gathered was not always integrated into the analysis.

The *idea* of beneficiary consultation is widely accepted and included in most terms of reference. It has however proved difficult to implement. Evaluation teams either fail to allocate enough time and resources or don't have relevant skills, or while acknowledging its importance find ways of evasion ('time did not permit' etc). Consultation with the affected population is one area where commissioning agencies should play a greater gatekeeping role, as well as ensuring that sufficient time has been allowed.

4.3.8 International standards

The conclusion from the evaluation reports and Chapter 3, in relation to international standards, is that their use remains somewhat limited, both in operations and in evaluations. While one third of the Kosovo reports and just under half of the non-Kosovo reports make reference to international standards, these are rarely integrated as evaluation tools to assess performance. Chapter 3 also notes the controversial nature of the standards, and the difficulties of establishing standards across regions and cultures. Evaluations need to make a distinction between assessing whether agencies have used international standards adequately in their work, and the use of standards to assess the intervention itself.

Evaluations that do integrate standards illustrate their use as evaluation tools and raise important evaluation questions (e.g., DANIDA, 1999c; DEC, 2000b, 2000c; UNHCR, 2000a). (UNHCR, 2000a) provides a field-level example through its use of international conventions and human rights law as a means of assessing UNHCR's performance in protection in the Kosovo conflict, and another is provided by the DEC's Orissa evaluation (DEC, 2000b). The latter used the Sphere Charter to open a dialogue with one NGO, and to evaluate its planned resource allocation to a marginalised group, contrary to the Sphere guidelines.

4.3.9 Attention to gender, the vulnerable or marginalised

Only about one third of the non-Kosovo evaluation reports contained findings on gender or could be considered to be partly gender-mainstreamed (Chapter 2). In the Kosovo reports gender is largely absent and one of the biggest gaps, remarkable in that a majority of adult refugees leaving Kosovo appear to have been women (information only imparted in passing). WFP notes that: 'Reports indicated the relocation of many women, children and elderly people from Pristina and southern towns in Kosovo to FYRoM to escape the fighting,' (WFP, 2000b: p33). The absence of attention to gender is even more remarkable given that, in determining that the intervention was both a success and

relevant to the needs of the affected population, almost nothing is said about its differential impact on women and men. Part of the reason for the more extensive gaps in the Kosovo evaluations may be that they tended to focus on political issues to the exclusion of cultural and social ones (see Chapter 3). The lack of adequate attention to gender is one of the most consistent gaps across both sets of evaluations.

In terms of attention to vulnerable or marginalised groups, the non-Kosovo set also fares much better than the Kosovo set with just over half of them covering this area in a satisfactory manner. In the Kosovo set only three paid sufficient attention to these groups. Many of the evaluations focused on geopolitics, internal politics and organisational issues, to the exclusion of a disaggregated analysis of the affected population. Many also display a top-down non-reflective approach in this area that is unlikely to support lesson learning. The views, perceptions and opinions of vulnerable and marginalised groups appear to have been largely ignored in a majority of evaluations, despite a moral obligation on the part of evaluators to represent the interests of this group, and the potential key information they can provide.

4.3.10 Coverage of factors potentially influencing performance

Reports were assessed as performing reasonably in terms of inclusion of coverage of factors such as geopolitics and the role of local authorities, although significant gaps remain, as pointed out in 4.3.5. Because many of the reports, and in particular the Kosovo set, had a focus on institutional performance, this is one of their strongest points. Many evaluators clearly felt more comfortable dealing with management issues such as hiring practices or head office/field communications, than they did trying to measure impact. This is perhaps not surprising because until recently, and in the last five years in particular, much of the evaluation of development cooperation has focused on intra-institutional issues. It has only been with the introduction of results-based management approaches in many agencies (see Chapter 1) that there has been an increased focus on measuring impact. Evaluations of humanitarian action appear to have some catching up to do in this area.

4.3.11 Conclusions and recommendations

Promoting dialogue about, and sharing the findings of, the draft evaluation has become common practice in the evaluation system. Of the 49 individual evaluations, 37 had adequately shared the draft at least with the commissioning agency, and a number of reports noted how they had responded to comments. This is a positive sign as it increases the likelihood that evaluation report recommendations will be acted on.

Both Kosovo and non-Kosovo reports were relatively strong in terms of the development of clear conclusions, and recommendations that were logically tied to conclusions, however, some were better than others. This is another area where the collection and dissemination of good practice material may be useful, for example where recommendations have been taken up by agencies. ALNAP's work on the follow-up to evaluations (seeEndnote 1, Chapter 1) should produce good practice examples in this area.

4.3.12 Legibility

Evaluators have recognised the importance of presenting evaluation findings in an attractive and easy to read format, and this has for the most part been done without sacrificing evaluation quality. Some, however, were less accessible and contained page after page of single-spaced text in small fonts, running in some instances to more than 150 pages. For the most part diagrams and visual aids were well used. Most reports contained an executive summary that accurately and concisely represented the

findings of the evaluation. Legibility may thus be seen as a comparative strength of this overall set of evaluation reports.

4.4 Implications of the LRRD Debate for Establishing Evaluation Standards

Given its particularly problematic nature in relation to the development of evaluation standards and criteria, the LRRD issue is raised again. As USAID notes: 'That relief experts have different views of the purpose of emergency assistance – whether it is for relief only, or for rehabilitation and economic development as well – exacerbates an already complex [methodological] situation,' (USAID, 2000a: p v). How an emergency or disaster is perceived and classified will have a major influence on which indicators of impact are used. While it is common in general evaluation practice to assess interventions against both stated intentions and a wider set of criteria such as impact and sustainability, some evaluation reports argue that these wider criteria are less relevant for the evaluation of humanitarian action (USAID, 2000; ECHO, 2000).

However, the OECD-DAC (1999) guidance for evaluation of complex emergencies includes the criteria of impact and sustainability/connectedness (see Endnote 3, Chapter 2, among the areas to be evaluated. It thus requires an evaluation against broader criteria than just intervention objectives, which it defines as effectiveness. In the general evaluation field, impact is usually considered a 'higher-level' indicator of results, and it may be that in the determining of the impact of humanitarian action, short-term and long-term objectives should be equally weighed. This would indicate a shift away from current practice, where short-term objectives (such as saving lives) are used as primary indicators. By asking fundamental questions about the results of an intervention, evaluation feedback mechanisms should help establish priorities and stimulate debate in this area.

The fact that many of the evaluations cover humanitarian action and rehabilitation, and in some cases development, has implications for evaluators and evaluation offices. Support documents (handbooks, good practice guidelines, etc) need to focus not only on crisis situations and relief, but also on how to evaluate the link to development.

4.5 Next Steps in Evaluation of Humanitarian Action

While there is much good practice in both Kosovo and non-Kosovo evaluation reports, and there appears to have been progress in some areas in terms of quality, much remains to be done. There is also a pressing need to learn from the wider evaluation field since the evaluations considered in this Annual Review have hardly drawn on lessons from this source. The quality of evaluations currently accepted by commissioning agencies is too low in several areas. There is a pressing need for the development of evaluation of humanitarian action standards, tools and capacity. These will help improve evaluation quality and credibility, as well as providing a means of holding evaluators accountable.

Development of standards and tools should adopt a dual focus, first in relation to the topic areas of impact, relevance, connectedness, etc, and second in relation to the 'new' agenda items such as protection and advocacy. Evaluators need clearer guidelines concerning what is expected from them in both areas, as well as the linkages between them. The relative weight to be assigned to each in order to determine the impact of the intervention also needs to be clarified. The majority of evaluations, whose purpose was to account for expenditure of material inputs, were perceived as apolitical. Yet, widening

their perspectives – i.e., factoring in aspects such as protection and human rights, both more politically sensitive – has system-wide implications in need of discussion. Can impartiality of evaluations be assured when assessing politically sensitive factors? The evaluation of protection also raises the issue of what expertise and experience there is in the analysis of protection in the evaluation community and, more particularly, in the case of evaluators from countries and regions primarily affected by complex emergencies and natural disasters.

The development of evaluation standards could follow the process that led up to the production of the US Program Evaluation Standards (Joint Committee, 1994), which, to the authors' knowledge, are currently the main set of standards in use in the wider evaluation field.

'The Joint Committee developed a systematic public process for establishing and testing the new standards before recommending their use ... the process involved many experts in evaluation, users of evaluation, and others ... Several hundred educators, social scientists, and lay citizens were involved in the development of the first program evaluation standards,' (ibid. p xvii).

Increased participation is likely to provide a wider range of experience and ensure buy-in. It is anticipated that the development of standards for the evaluation of humanitarian action might be a two- to three-year process at the least.

'Standard A4', an example which illustrates the approach of the 30 Program Standards, covers 'Defensible Information Sources' as follows: 'The sources of information used in a program evaluation should be described in enough detail, so that the adequacy of the information can be assessed,' (ibid, p141). Clearly, the quote under **Consultation with beneficiaries** ('In the field the team spent as much time as possible with the beneficiaries,') would not meet the standard. The Joint Committee (1994) also provides an overview of the standards, guidelines, common errors and illustrative cases.

Specific tools such as the proforma used in this Review can be developed in parallel, again in a participatory fashion, and could include tools for the assessment of protection. Good practice material could also be disseminated on, for example, the provision of contextual information and the crafting of recommendations.

In the meantime immediate steps could be taken, with minimal effort and cost, to increase the transparency and credibility of evaluations, all related to the provision of more information:

- the commissioning agency can note in an introduction how the evaluation team was selected, and the intended circulation and use of the evaluation;
- the evaluation can note constraints and show how these have impacted on data gathering and analysis;
- the evaluation can provide details of the methodology, and in particular disaggregated information on consultation with the affected population, including numbers consulted, and methods used for consultation and analysis of data;
- the evaluation can provide a short biography of team members, noting their regional and sectoral expertise, any work in the gender field, their language capacity and their experience in evaluation (e.g., DEC, 2000);
- all data should be gender-disaggregated where possible.

Finally, there were instances where the same agency had almost simultaneously commissioned evaluations where one had been strong and the other weak in exactly the same area. This suggests commissioning agencies can improve on their role as gatekeepers of evaluation standards and practice, particularly in relation to identified gaps.

4.6 Concluding Remarks

This Annual Review has demonstrated the inherent value of bringing together and systematically synthesising the findings of a large set of evaluations of humanitarian action. By highlighting that these are system-wide performance issues, whether areas found wanting or good practice deserving emulation, it has confirmed their relevance to all within the international humanitarian system. Its principal findings should, therefore, help set the agenda for ongoing discussions on how to improve performance.

The quality assessment of the evaluation reports, through meta-evaluation, also reveals a substantial agenda for those involved in the commissioning, managing, undertaking and use of evaluations of humanitarian action.

The Review makes a number of criticisms of current evaluation practice and standards in some of the more important and substantive areas, including methodological rigour and transparency. This is done in the context of the need to improve evaluation practice, as evaluation and results-based management are here to stay for the medium-term future at least (if only due to the requirement that agencies illustrate adequate results to funders and governments).

The precise form of evaluation of humanitarian action is still under development, and some have questioned whether the evaluation mechanism needs to be complemented by social audits (Raynard, 2000). It is clear, however, that if sufficiently rigorous and adherent to good practice standards, evaluation of humanitarian action has the potential to make a significant contribution both to accountability and to lesson learning, leading to improved practice.

Endnotes

1. Work to develop criteria for the evaluation of protection activities which draws on those so far developed by UNHCR and ICRC is planned by ALNAP as part of its Evaluation 'Guidance Gap Filling' activities.
2. Information on the Network is available on the Global Peacebuilding Network site at <http://wbln0018.worldbank.org/ESSD/pc1.nsf/Home>
3. Information on People in Aid is available at www.peopleinaid.org
4. These figures are necessarily approximate given the aggregate nature of the assessments.
5. In the current Review there are clear biases, e.g., in terms of attention to protection, and in the choice of topics for the proforma. Further biases are spelt out in Chapter 3.
6. Some evaluations (e.g., ICRC/IFRC, 2000; ECHO, 2000p) used local research teams from the agency being assessed, but the involvement of these teams, or the implications of this for the evaluation, is not noted in detail.
7. The Joint Committee on Standards for Educational Evaluation (Joint Committee, 1994: p133) gives as the standard for 'Context Analysis': 'The context in which the program exists should be examined in enough detail, so that its likely influences on the program can be identified.'
8. This is in contrast to the long lists of agency staff consulted that usually made up an Annex to the evaluations.