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 12 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 co-operation 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 with sufficient depth and 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 (Chapter 1, Endnote 1).

4.2.15 Humanitarian Evacuation and Transfer Programmes

The Humanitarian Evacuation and Transfer Programmes 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 – e.g., 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 the evaluation. 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 in terms of information provided on the nature and objectives of the intervention, as well as clarity of purpose. 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 one by host country citizens. Of the 55 evaluators involved in the Kosovo evaluations, 52 were expatriates, and 53 were independent consultants (with two 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 actually takes place in a meaningful fashion. 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 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 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, and evaluate one NGO's 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 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 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 the section on Information on Context above. 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 used.

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 (see Chapter 1 endnote 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 over 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 Chapter 2, Endnote 3) 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 imply a shift 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 to hold 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 will likely 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.

An example to illustrate the approach of the Program Standards – encompassing 30 standards. 'Standard A4' 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 Annual Review makes a number of criticisms of current evaluation practice and standards in some of the more important and substantive areas, such as 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 to both accountability and lesson learning, leading to improved practice.

ENDNOTES

Chapter 1

1 As well as the Annual Review series, ALNAP's 2001-02 Workplan activities in relation to the evaluation of humanitarian action include: the publication of a volume of reflections by evaluation practitioners on nine humanitarian action evaluations undertaken between 1993-2000 (Wood et al, 2001); the development of short courses and training modules for agency personnel and evaluators, studies on how evaluations of humanitarian action are used and followed-up by agencies, and, the sharing of evaluation plans between members to reduce duplication and encourage greater collaboration in evaluation. Other accountability and learning activities include a book on accountability frameworks in the international humanitarian system; a major international collaborative study to develop guidance for humanitarian agencies on consultation with and participation of beneficiaries and the affected population; the testing the field level 'Learning Support Office' (see Endnote Chapter 3, 5)

2 The ALNAP Symposium 'Learning-from-Evaluation: Humanitarian Assistance and Protection in Kosovo' was held in Geneva on 17th & 18th October 2000. Some 51 representatives (Military/Red Cross Movement / NGO/Bilateral and Multilateral Donor Organisations/UN plus Academics and Consultants) from 43 different organisations that had been involved in the Kosovo crisis or the evaluation of humanitarian programmes, participated in the event

3 Operation Provide Comfort was mounted in the aftermath of the Gulf War between Coalition Forces and the Iraqi forces that had occupied Kuwait. Following the March 1991 ceasefire between the Coalition Forces and the Iraqi Government, a popular uprising took place in the predominantly Kurdish areas of northern and eastern Iraq The brutal suppression of the uprising by Iraqi military led to the exodus of almost 2 million Kurds towards Iran and Turkey. The Turkish authorities refused them admission and almost 500,000 Kurds were trapped in the mountains along the Iraqi/ Turkish border. Following Security Council Resolution 688 of 5 April, US and European airforces began airdropping supplies and US, British and French forces created a 'Safe Haven' in northern Iraq enabling the provision of direct humanitarian assistance by the militaries, the UN, the Red Cross and NGOs and the movement of the Kurds down from the mountains back into Iraq The operation was a watershed event for the humanitarian system as it set a precedent for military intervention in support of humanitarian objectives on sovereign territory that was subsequently repeated in Somalia, Kosovo and East Timor

4. ICRC, IFRC, UNICEF, UNHCR and WFP are the 5 organisations involved in the benchmarking study

5. The DAC Working Party on Aid Evaluation maintains a database of evaluations of development cooperation undertaken by bilateral and multilateral donor organisations <www oecd.org/dac/evaluation>. Whilst the database is extremely valuable, evaluations undertaken by organisations other than DAC members, in the international humanitarian sector are not included and the number of evaluations of humanitarian action is limited.

6. see Chapter 2 Endnote 4.

7. The book is scheduled for publication by Zed Press in August 2001.

8. The proforma has been given a blind pre-test by two independent consultants to assess consistency of application across different assessors and help guide the proforma's further development/refinement. The objective is to strengthen its methodological basis to ensure its acceptance

Chapter 2

1. Use has also been made of a paper (HPN, 2000) that compares the approaches and results of an evaluation included in the set (DEC, 2000) with two evaluative studies of humanitarian operations in the same region, not included in the set (Groupe URD, 2000; Lister, 2000).

2. The DANIDA evaluation reports cover the period 1992-99, otherwise most reports cover (1999 and 2000).

3. Connectedness concerns the need for humanitarian programmes to take longer term needs into account in addition to their focus on addressing immediate, acute needs. The criteria is similar to the standard evaluation criteria of 'sustainability' but takes account of the reality that the majority of humanitarian programmes are not intended to be sustainable (OECD-DAC, 1999).

4. Triangulation is a key technique allowing cross-referencing and cross-checking of different information sources. In complex emergencies/ disasters where data may in general be poor, triangulation will support evaluation rigour.

5. Part of UNHCR's response, to the ECHO (2000b) evaluation of support to rehabilitation programmes in Rwanda, was to note that the evaluation had paid insufficient attention to the protection aspects of the intervention.

6. HPN and ALNAP literature was widely used in the evaluations, both for the development of methodology and for comparative purposes.

7 The question of 'impartiality' is under discussion in the general evaluation field at present, where there has been a questioning of the concept of the evaluator as an objective and unbiased observer. Related to this is the concept that evaluations, to be effective and used, need to be as participatory as possible.

8 The latter two figures are estimates based on the schedules in the reports.

9. A control group is a group chosen randomly from the same population as the group that benefited from the programme but for some reason the control group did not benefit from the programme. A control group therefore enables an assessment to be made of what would probably have happened to the programme group if the programme had not been undertaken.

10. In one report the area was not applicable.

11. Interview with Ed Tsui, Director, Policy, Advocacy and Information Division, UNOCHA, February 2001.

Chapter 3

1. A much shorter, but structurally similar, form of this paper was presented and discussed at the ALNAP symposium 'Learning from Evaluation: Humanitarian Assistance and Protection in Kosovo', which took place in Geneva from 17–18 October 2000. The title of that paper was 'Kosovo Humanitarian Programme Evaluations: Towards Synthesis, Meta-evaluations and Sixteen Propositions for Discussion'. It is now presented, in greater depth as Chapter 3. This takes into account discussions from the plenary sessions which focussed on building immediate response capacity, improving civil military cooperation, achieving social learning, and reconceptualising humanitarian programme evaluation as critical learning practice. Evaluations not available earlier were obtained and considered.

   It is important to emphasise that the purpose of this paper is to provide an overview of common (and dissenting) themes and messages, not highlight one evaluation or study over another.

2. see Chapter 1 Endnote 2

3. See (Scriven, 1991)

4. See (Pottier, 1999)

5. The objective of the ALNAP Learning Support Office concept is to make a positive impact on the quality of emergency response in the field through the promotion and facilitation of three-way learning activities: i. 'learning-

in', ii 'learning-out', and iii 'lateral learning' It is proposed that a test Learning Support Office will be run in Sierra Leone in 2001.

Chapter 4

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, 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.

ANNEXES

ANNEX 1

ALNAP PROFORMA FOR USE IN ASSESSING THE QUALITY OF EVALUATION REPORTS OF HUMANITARIAN ACTIONS

This 'quality proforma' is an initial attempt to provide the humanitarian system with a tool by which to measure the quality of its evaluation reports. It draws on current evaluation guidance and a growing body of what is increasingly acknowledged as good practice in evaluation. Sources include: Valadez and Bamberger (1994); Patton (1997); OECD-DAC (1999); Apthorpe (2000); Raynard (2000); Sphere (2000); Wood, et al. (2001). For definitions see OECD-DAC (1999).

Non-rated	*Non-rated*
1. Purpose and Focus of the evaluation	**2. Constraints**
i. Motivating factor for the evaluation (if stated)? *e.g., donor requirement, agency commitment to learning* ii Primary purpose of the evaluation? *e.g. accountability, lesson-learning?* iii If multipurpose, what was the relative emphasis? iv What was the primary focus of the study? *e.g., partner performance, programme, project* v Was a mixed (internal and external) or external evaluation team used? vi Is the report in the public domain?	i Stated constraints that substantively hindered the evaluators' ability to meet the evaluation's purpose and/or adhere to good evaluation practice? *e.g., ability to visit desired areas, access to relevant documentation, access to interviewees*
Assessor comments	
N.B The 'Assessor comments' boxes may be used to link a low rating to stated constraints (Column 2) as well as for general points of clarification	

Rating system to be applied on the basis of evidence contained, or not, in the report **A** none provided/no evidence, **B** poor, **C** satisfactory, **D** good, **E** very good; **F** excellent

Rated	*Rated*	*Rated*
3. TOR, team composition and time allowed	**4. Information on context and intervention**	**5. Methodology and transparency**
☐ i Quality of statement in the TOR on the nature, objectives stakeholders of the intervention to be evaluated?	☐ i Quality of contextual information provided on the affected area, the factors contributing to the crisis and (if relevant) its continuation?	☐ i Quality of statement of overall evaluation approach and methodology? *e g , did it extend to 'methods used'*
☐ ii Clarity of purpose and focus of evaluation as outlined in the TOR? *e g , are the purpose/focus consistently addressed/reflected throughout?*	☐ ii Provision and quality of 'narrative history baselines'? *e g , did it take account of views of relevant actors?*	☐ ii Use made of good evaluation practice techniques? *e g , multi-method approach, triangulation*
☐ iii Awareness of the role of policy on the intervention to be evaluated, demonstrated in the TOR?	☐ iii Use made of contextual information in analysis?	☐ iii Use made of generally accepted evaluation criteria? *e g , efficiency, effectiveness, impact, relevance, connectedness, coherence, coverage?*
☐ iv Awareness of longer-term perspectives demonstrated in TOR? *e g , impact of intervention on future risk*	☐ iv Use made of comparative perspectives? *e g , through reference to, and use of, other similar evaluations*	
☐ v Appropriateness of team composition? *e g , did it allow for specialist inputs in areas to be covered, including local knowledge expertise?*	☐ v Verification of the objectives of the intervention, as stated in TOR or original Log Frame, in respect of relevance and variation?	
☐ vi Appropriateness of time allowed for the evaluation?	☐ vi Use of data sources, including secondary documentation?	
☐ vii Involvement of evaluation team in finalising the TOR?		
Assessor comments		

Rating system to be applied on the basis of evidence contained, or not, in the report. **A** none provided/no evidence, **B** poor, **C** satisfactory, **D** good; **E** very good, **F** excellent.

Rated	*Rated*	*Rated*
6 Consultation with *Beneficiaries* & Affected Population	**7. Reference to *international* standards**	**8. Attention to *gender and the* vulnerable or marginalised**
☐ i Quality of consultation undertaken with a representative sample of relevant beneficiaries within the affected population?	☐ i Appropriate reference to international standards? *e g , relevant areas of international humanitarian and human rights law, the Red Cross/ NGO Code of Conduct, and developing standards such as Sphere?*	☐ i Consideration given to gender issues? *e g are these visibly and meaningfully addressed in the report*
☐ ii Quality of consultation undertaken with a representative sample of relevant non-beneficiaries within the affected population?		☐ ii Consideration given to needs of vulnerable or marginalised groups within the affected population? *e g children, the elderly, disabled, HIV and Aids sufferers Are these visibly and meaningfully addressed?*
☐ iii Use made generally of the material gained through consultation? *e g , assessment of targeting re those most in need*		
Assessor comments		