Meta-evaluation

Peter Wiles

4.1 Introduction

The purpose of the annual ALNAP meta-evaluation is to provide an overall assessment and review of evaluations of humanitarian action (EHAs) in order to identify trends in quality. This chapter also aims to provide examples of good practice, and to capture some of the dynamics of EHA, arising from feedback from evaluation managers and evaluators. The main aim of the meta-evaluation is to provide information for improvements in practice. As defined by Lipsey (2000), meta-evaluation is 'meta-analysis and other forms of systematic synthesis of evaluations providing the information resources for a continuous improvement of evaluation practice'.

This meta-evaluation is based on an assessment of 30 evaluation reports on evaluations conducted during 2003 and 2004, and submitted to ALNAP for this year's Review of Humanitarian Action[1] (Annex 3). Section 4.2 below covers the meta-evaluation process, while Section 4.3 gives an overview of the reports in this year's set. Section 4.4 focuses on joint-agency initiatives, and Section 4.5 provides an overview of the meta-evaluation results. Section 4.6 looks in detail at each of the five main elements of the quality proforma (terms of reference; evaluation approach, methods and practice; contextual analysis; the intervention; the report). Section 4.7 deals with overall conclusions, and Section 4.8 looks at future issues for the ALNAP meta-evaluation.[2]

A full listing of this year's results has been included as annex 1, while the revised Quality Proforma and full list of source reports submitted for review by agencies have been included as annexes 2 and 3 respectively. The chapter also attempts to make some comparisons with the results of previous years as a means to identify trends in EHA practice.

4.2 The meta-evaluation process

The meta-evaluation process uses the ALNAP quality proforma (QP) as the tool for rating the quality of humanitarian evaluations and tracking trends over time. First developed in 2000, the quality proforma draws on commonly accepted good practice in EHA and evaluation in general. The proforma is also intended to be used as a checklist for evaluation practitioners, both managers and evaluators. It is included in full as Annex 2. The quality proforma has been developed and adapted over the five years of its use. In 2004, in consultation with a group of ALNAP members, it was thoroughly reviewed, primarily to make it easier and clearer to use, as well as making it slightly shorter.[3] Additional guidance notes and a glossary were also included in the quality proforma.[4]

The evaluation reports were analysed by two assessors with a total of 45 years in international humanitarian and development work, each with over 10 years of humanitarian evaluation experience.[5] As in previous years, the assessors worked separately, and then together discussed major differences in ratings. At the end of this iterative process, there was 83 per cent agreement between the assessors.[6]

Historically, the core of the ALNAP meta-evaluation has been the rating of the selected written reports using the quality proforma. However, a report that is overall rated as poor by the assessors, may, in terms of outcomes, have proved very useful to the commissioning agency and to other key stakeholders, or *vice versa*. For this reason, as in the last two years, in addition to the desk assessment exercise using the quality proforma, 11 of the 18 commissioning agencies were contacted and evaluation managers interviewed to gain a better understanding of the processes surrounding the evaluations. This process enabled follow-up and discussion to be carried out on two-thirds of the reports in this year's evaluation set.

This year, for the first time, evaluators were also brought into the meta-evaluation process: 13 evaluators were contacted by questionnaire, and 7 replied. In all but one case, evaluators and managers were both consulted about their respective reports, thus giving a rounded view of the evaluation processes. The response from evaluators to this initiative was generally positive. One wrote: 'This process has been extremely useful for me personally, to see how others rate my work, so that I can identify areas that need to be improved.' This should be seen in a context where

about 50 per cent of the evaluators said that they did not receive any specific feedback from the commissioning agency about the quality of their work.

The interactions with evaluation practitioners also provided the opportunity for them to comment on the ratings given to their reports by the ALNAP assessors. Feedback to the ALNAP secretariat on the meta-evaluation process was also gained. Both managers and evaluators were asked to list their top three factors that were most likely to result in a good evaluation. The results of this small survey are included below in Section 4.7.

# 4.3 Overview of the evaluation reports

The 30 reports in the meta-evaluation were selected from the ALNAP Evaluation Reports Database (ERD), to cover a mixture of commissioning agencies, including donors, UN agencies and NGOs. Commissioning agencies give consent to have their reports included in the meta-evaluation process. Reports are selected for the meta-evaluation to fit with ALNAP's definition of evaluation of humanitarian action:

> A systematic and impartial examination of humanitarian action intended to draw lessons to improve policy and practice, and enhance accountability. It has the following characteristics: i) it is commissioned by or in cooperation with the organisation(s) whose performance is being evaluated; ii) it is undertaken either by a team of non-employees (external) or by a mixed team of non-employees (external) and employees (internal) from the commissioning organisation and/or the organisation being evaluated; iii) it assesses policy and/or practice against recognised criteria (eg efficiency, effectiveness/timeliness/coordination, impact, connectedness, relevance/ appropriateness, coverage, coherence and as appropriate, protection); and iv) it articulates findings, draws conclusions and makes recommendations. (ALNAP, 2001, p 18)

The meta-evaluation, therefore, does not select reports from evaluative exercises such as internal reviews or lessons-learning exercises that are not considered full

evaluations. Internal agency documents are not always included in the ALNAP database. Increasing decentralisation of donors and other agencies may mean that more evaluative work is commissioned and carried out at regional and country level, with reports not necessarily reaching headquarters. This raises the question of whether ALNAP should seek to increase its profile and contacts beyond agency headquarters.

The ALNAP quality proforma was designed to assess evaluation reports dealing with humanitarian responses. It is less well suited to assessing evaluations of an institutional or policy nature. For this reason, such reports may not be included in the meta-evaluation. For example, the external evaluation of OCHA's Internal Displacement Unit was not included in this year's evaluation set. Figure 4.1 shows the origin of the reports chosen for this year's meta-evaluation.

Figure 4.1 Origin of evaluation reports according to type of agency
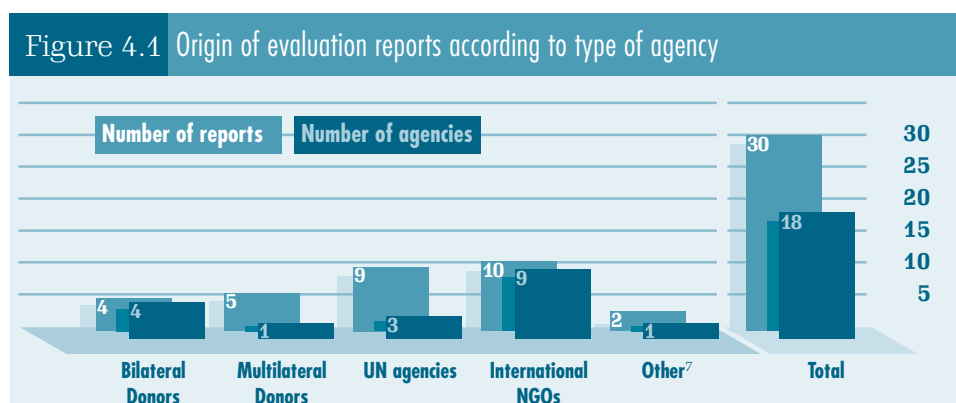
As in previous years, ECHO and WFP have been major contributors to the set, with five reports from each of these agencies.[8] Of the international NGOs, Tearfund UK submitted three reports. Therefore, ECHO, Tearfund and WFP together contributed just under half (43 per cent) of the report set. Two-thirds of the reports evaluated work in Africa, with most of the remainder in Asia.[9]

Not all ALNAP member agencies submitted evaluation reports to the Secretariat for the 2004/2005 meta-evaluation. It is therefore important to note that the reports selected for the meta-evaluation are not a representative sample of the EHA output

from the humanitarian sector. For example, there are no reports from the International Federation of the Red Cross and Red Crescent or the ICRC in this year's selection of reports.

Evaluation reports commissioned by national governments or national NGOs rarely appear in the report sets for the meta-evaluation. This probably reflects the composition of ALNAP's membership, being primarily Northern donor and aid agencies. An exception in this year's set is the evaluation of the response to the 2002–03 emergency in Ethiopia that was jointly commissioned by the Ethiopian government's Disaster Prevention and Preparedness Commission, along with the UN Secretary General's Special Envoy for the Horn of Africa and the UN Strategic Disaster Management Team. Also in the NGO sector, an evaluation of a USAID-funded seed programme in Tigray, Ethiopia, was jointly commissioned by a national and an international NGO.

Many of the reports in this meta-evaluation set deal with chronic long-term emergencies such as those in Zimbabwe and the Democratic Republic of Congo, or refugee health in long-term camp situations. A small number of reports cover slow-onset natural disasters, such as drought in Ethiopia, and there are no cases in this set of sudden-onset disasters. As a result, in contrast with the 2003 meta-evaluation set, there are no real-time evaluations in this year's group of reports.[10]

4.4 Joint evaluations

Six reports in this set are joint-agency evaluations, reflecting a trend towards this approach in evaluative activity. Three reports, the pilot phase of a programme to establish inter-agency health programme evaluations in humanitarian crises, are joint UNHCR / WHO evaluations looking at refugees in camps in Nepal, Pakistan and Zambia (Box 4.4). The concept paper for these three evaluations noted that:

Many agencies have monitoring and evaluation systems in place but these usually cover only a limited range of activities within the health sector. While together all agencies have a responsibility toward the entire affected

population and the sector wide response they receive, there has been little analysis done to review whether the efforts of all health agencies combined are making a coherent, effective and appropriate contribution to achieve the best possible health outcomes. (Griekspoor and Burns, 2003, p 1).

This impetus was summarised as 'Evaluating together what we cannot evaluate alone'.[11]

Box 4.1	Inter-agency health evaluations in humanitarian crises — lessons learned

- The scope of inter-agency evaluations should be sector-wide and aimed mainly at policy issues.

- While accountability issues will be looked at, the primary focus of inter-agency evaluations should be lesson-learning to improve performance.

- Evaluation is inextricably linked to both assessments and monitoring and therefore it is important to support better-integrated and harmonised systems of assessment, monitoring and evaluation.

- Development of terms of reference would benefit from more attention to process, and more consultation.

- Full agency buy-in is important – not just at headquarters but also in the country offices.

- Early involvement of national governments is also important, and inter-agency health evaluations should preferably be managed by in-country coordinating bodies of health sectors.

- Inter-agency evaluations should have a pragmatic focus and provide better information on which to base decisions, as opposed to perfect information.

- To ensure the consistent quality of inter-agency evaluations, analytical frameworks, standards and indicators to be used should be agreed and incorporated into generic terms of reference.

Sources WHO/London School of Hygiene and Tropical Medicine/UNHCR, 2004; author's interviews.

Three joint-agency reports dealing with assistance to IDPs in Angola, the Democratic Republic of Congo and Indonesia are part of a global IDP evaluation initiative started in May 2003 by the Danish and Netherlands Ministries of Foreign Affairs, Sida and ECHO, subsequently joined by other agencies.[12] The objective of this joint-agency process was to draw out key, system-wide lessons and thereby to improve the provision of humanitarian assistance and protection to IDPs in the future (Box 4.2).

Box 4.2	Inter-agency evaluations of support to IDPs – approaches and lessons learned

This initiative was based on a synthesis of findings from 17 evaluation reports. The foundation for the approach to this work was the *Framework for a Common Approach to Evaluating Assistance to IDPs* (Danida, 2003), which identified key issues at both policy and operational level that needed to be looked at, and relevant questions to be addressed in order to evaluate overall impact and effectiveness of assistance to IDPs using the standard DAC criteria.

The synthesis report (Borton et al, 2005) makes the following points about the experience of carrying out this work:

■ While the common framework gave a measure of commonality of approach and method for 6 of the 17 studies, the rest of the set, for a variety of reasons, did not follow the framework. The synthesis team therefore supplemented the evaluations with additional sources of information to maximise the learning potential of this collaborative exercise.

■ The material in the reports was analysed against 45 'key terms' derived from the terms of reference.

■ The process was seen as an attempt by a group of interested organisations to pilot a new, lighter approach to collaborative evaluation, which achieves many of the same benefits in terms of system-wide learning and policy development as those achieved by the Joint Evaluation of Emergency Assistance to Rwanda (JEEAR).

■ Overall, it was found that the collaborative exercise had proved popular with participating donor organisations.

■ The consultancy team felt that the approach should be repeated for other selected themes, but that it should be undertaken with more rigour and discipline – especially in relation to the common framework – and conducted over a longer period.

The most notable joint-agency, system-wide humanitarian evaluation carried out to date was the Joint Evaluation of Emergency Assistance to Rwanda (JEEAR) in 1996. In the development context, the OECD/DAC Network on Development Evaluation[13] has been promoting joint evaluations as a tool towards increased rationalisation of the process of evaluation, aiming to reduce transaction costs for partner countries, and to improve the quality of the work undertaken and the legitimacy of the evaluation (OECD/DAC, 2000; Dabelstein, 2003; Feinstein and Ingram, 2003).[14] Further impetus for donors to collaborate in their evaluation work has been provided by the Good Humanitarian Donorship initiative.

By the end of 2005 there should be considerably more experience and information amassed about joint evaluations. Further inter-agency health evaluations will have been completed, and the planned OECD/DAC publications on lessons from joint evaluations should be available (OECD/DAC, 2005).

4.5 Overall results of the meta-evaluation

The overall quality of evaluation reports this year remains unimpressive. Figure 4.2 gives the overall ratings for the five main areas of the quality proforma (QP). It shows that on average only just over half of the reports are rated as satisfactory or good. The weakest overall areas remain the terms of reference, and the sections of the reports dealing with evaluation methods, practice and constraints. As with all sections of the quality proforma, there is considerable variation in results within those broad areas. These variations are analysed in the following sections of this chapter.

Differentiating meta-evaluation results by type of commissioning agency, the results are clustered around between 50 and 54 per cent satisfactory or good ratings, with donors marginally best, followed by NGOs and then UN agencies (Figure 4.3). A larger sample of agencies over a period of years would be needed to make firmer conclusions about these figures. However, one can note that the quality of inter-national NGOs' reports appears comparable to those of donors and UN agencies.
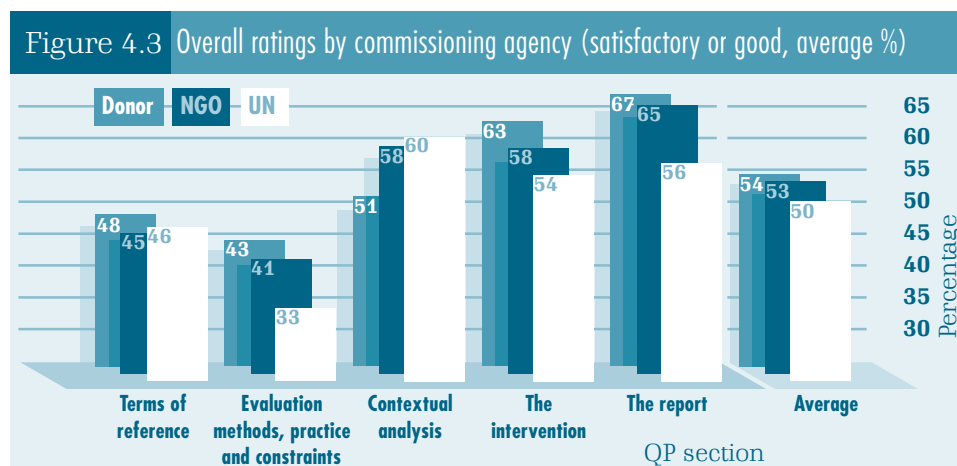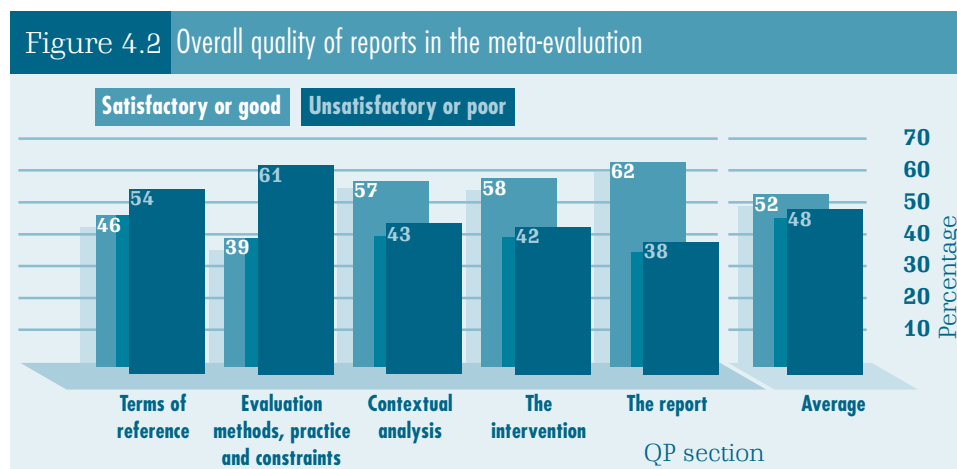
Figure 4.2 Overall quality of reports in the meta-evaluation

Satisfactory or good | Unsatisfactory or poor

Percentage

Terms of reference: 46, 54
Evaluation methods, practice and constraints: 39, 61
Contextual analysis: 57, 43
The intervention: 58, 42
The report: 62, 38
Average: 52, 48

QP section

Figure 4.3 Overall ratings by commissioning agency (satisfactory or good, average %)

Donor | NGO | UN

Percentage

Terms of reference: 48, 45, 46
Evaluation methods, practice and constraints: 43, 41, 33
Contextual analysis: 51, 58, 60
The intervention: 63, 58, 54
The report: 67, 65, 56
Average: 54, 53, 50

QP section

Table 4.1 Overall ratings by commissioning agency (satisfactory or good, average %)

	The ToR should clearly describe:	Satisfactory or good (%)	Unsatisfactory or poor (%)
1.1a	the work to be evaluated including its objectives and key stakeholders	46	54
1.1b	the purpose, objectives and focus of the evaluation	87	13
1.1c	the intended use and users of the evaluation outputs and responsibility for follow-up	40	60
1.1d	desired report framework	37	63
1.1e	rationale for timing of the evaluation	54	46
1.1f	evaluator selection process (eg bidding)	21	79
1.2	The ToR should clarify the commissioning agency's expectation of good humanitarian evaluation practice	37	63
	Average	46	54

4.6 Detailed results of the meta-evaluation, using the quality proforma[15]

This section covers the five main areas of the quality proforma, analysing overall ratings and trends, and highlighting good practice.

4.6.1 Terms of reference

Feedback gathered as part of this meta-evaluation indicates that both evaluation managers and evaluators rate the quality of the terms of reference of an evaluation as one of the most important factors in the EHA process. Good, realistic terms of reference with clearly stated objectives, adequate budget and realistic timeframe are all seen as crucial issues. It is therefore worrying that the average of satisfactory or good ratings for the terms of reference (ToR) quality proforma criterion was only 46 per cent (Table 4.1).

On the positive side, a substantial majority of terms of reference (87 per cent) clearly described the purpose, objectives and focus of the evaluations. Only just over one third of terms of reference (37 per cent) clarified the commissioning agency's expectation of good humanitarian evaluation practice, to a satisfactory or good standard. However, this is a major improvement on the two previous years' average of just 5 per cent for the corresponding value (Chart 4.1).

The terms of reference for the Terre des Hommes Foundation evaluation of an integrated Maternal and Child Health and Psycho-social Programme for the Eastern Province of Sri Lanka was one of the few to summarise clearly the objectives of the programme to be evaluated, extracting the information from the programme log frame. The terms of reference also concisely described the scope of the evaluation, its specific objectives and the methodology.

Evaluation buy-in and follow-up

There appears to be an increasing emphasis by evaluation managers on the importance of getting internal buy-in for evaluations, reflecting a continued concern

about the usability of evaluation reports and the outcomes of the investments made. There is a concern to ensure that staff members are clear about the purpose of the evaluation, and what they want to get out of it. In some agencies, including WFP, there is a move away from generic terms of reference so that the felt needs of country offices can be taken into account in order to ensure buy-in. It has been felt that sticking to generic terms of reference could jeopardise staff buy-in.
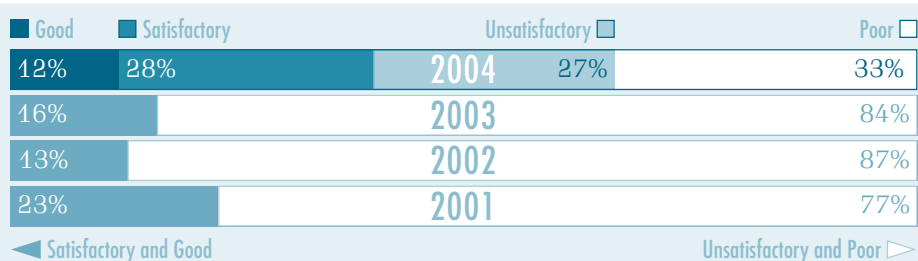
Chart 4.1	Results for QP Area 1.2, expectation of good practice

The terms of reference should clarify the commissioning agency's expectation of good humanitarian evaluation practice (eg application of DAC criteria; reference to international standards including international law; multi-method approach, ie, quantitative and qualitative; consultation with key stakeholders to inform findings, conclusions and recommendations; and gender analysis).

■ Good ■ Satisfactory	Year review[16]	Unsatisfactory ■	Poor ☐
6% 31%	2004	38%	25%
7%	2003		93%
3 %	2002		97%
40%	2001		60%

◄ Satisfactory and Good   Unsatisfactory and Poor ▷

Chart 4.2	Results for QP Area 1.1c, terms of reference – Intended use and users of the evaluation

The terms of reference should clarify intended use and users of the evaluation outputs and the individual or department responsible for follow-up.

■ Good ■ Satisfactory		Unsatisfactory ■	Poor ☐
12% 28%	2004	27%	33%
16%	2003		84%
13%	2002		87%
23%	2001		77%

◄ Satisfactory and Good   Unsatisfactory and Poor ▷

| Box 4.3 | Good practice in drawing up terms of reference |

The terms of reference of the Disasters Emergency Committee (DEC) continue to demonstrate good practice. For the evaluation of the Liberia Crisis Appeal, the terms of reference included the following features.

■ A concise overview of the emergency context.

■ An overview of the DEC member agencies' financial commitments.

■ A clear statement of the DEC's evaluation policy, noting that there is a tension between evaluating the impact and strategic use of the DEC funds as a whole, and the performance of the individual member agencies: '[the] evaluation should review the overall impact of the Members but a real effort should be made to root these conclusions in a substantial number of… individual member's performance'.

■ Expectation of the criteria, frameworks and methodology to be used by the evaluator, including matching the DAC criteria with the Red Cross NGO Code of Conduct, to which all DEC members sign up. A detailed annex lists evaluation questions and indicators to be used under each of the Principles of Code of Conducts.

■ The terms of reference for the Liberia appeal build on lessons from earlier DEC evaluations and highlight issues to which the evaluator needs to pay special attention:

  • scaling up and phasing out;
  • gender analysis and differentiation in programme activities;
  • protection;
  • accountability to beneficiaries.

Debates in this area highlight the continuing tension between the accountability and lesson-learning objectives of most evaluations. Is there a danger that staff concerns about the practical problems and about the future direction of programmes will dilute the strict accountability functions, which may of course highlight their own performance? This focus on buy-in and usability may be borne out by an improvement in the terms of reference descriptions of the intended use and users of the evaluation (Chart 4.2), albeit starting from a low base.

The terms of reference for WFP evaluation of the Eritrea Relief Portfolio (WFP, January 2004) is one of the few terms of reference to describe internal accountability for the evaluation, and how the evaluation products and processes will be used:

- **formative end use** by the WFP country office and the Eritrean government for supporting and encouraging ongoing programme modifications and improvements and for formulating future interventions;

- **summative end use** by the WFP evaluation office to provide accountability to the WFP executive board;

- **conceptual end use** by WFP corporately for identifying generic lessons and for encouraging group and staff learning.

A management response matrix is prepared by the WFP evaluation office and the country office, to track actions supporting follow-up to an evaluation.

Realistic terms of reference

While about half of the evaluation managers interviewed felt that the terms of reference were realistic, a significant number felt that they were in danger of becoming overloaded with too many questions. The terms of reference for one evaluation in this year's set included over 50 questions to the evaluators, some of which were very complex. The report's authors noted that the issues to be covered were excessive and unprioritised. It was agreed by the commissioning agency that the consultants had discretion to prioritise issues themselves in the field, according to feasibility and perceived importance. The evaluators decided to cover a broad range of issues, but not in great depth.

However, at the end of the evaluations overall, most managers were satisfied that the reports had met the requirements of the terms of reference. A majority of evaluators thought that the terms of reference given were just about realistic, with a minority feeling that they were overloaded.

Risk assessment

Terms of reference rarely include a section on risks, and this is not specifically included in the quality proforma. However, particularly in the context of conflict

situations, it seems good practice to mention and attempt to assess the risks that might compromise an evaluation, for example in terms of lack of access to certain geographical areas. The Sida evaluation of IDPs in Indonesia included a section in the terms of reference weighing up risks of non-access to areas for reasons of security and possible natural disaster. In fact, a change was made in the choice of case study locations because of renewed conflict in one area, and Aceh was not visited by the full team. The commissioning agency felt that these security issues did not affect the quality of the report.

4.6.2 Assessing evaluation methods, practice and constraints

As shown in Figure 4.2 above, this section has the lowest average quality of all the sections in the quality proforma. The ratings for each criterion are given in Table 4.2, showing that an average of only 39 per cent of reports have satisfactory or good ratings here. This somewhat gloomy picture is offset by the fact that there has been a significant improvement in the important area of description of evaluation methods and their appropriateness to the evaluation's main aim (Chart 4.3). Without this information, it is hard to verify the validity of a report's methodology.

Table 4.2	Results summary for QP Section 2, assessing evaluation methods, practice and constraints		
		Satisfactory or good (%)	Unsatisfactory or poor (%)
2.1a	Nature and make up of the evaluation team	25	75
2.1b	Outline of evaluator biases that might have affected the evaluation	9	91
2.2	Any clarification process between the agency and the evaluators on the scope and methods of the evaluation that resulted in modifications to the ToR	31	69
2.3	Appropriateness of overall evaluation methods	59	41
2.4a	Consultation with and participation by primary stakeholders	48	52
2.4b	Consultation with other key stakeholders	57	43
2.5	Assessment of intervention against appropriate international standards	28	72
2.6	Constraints on the evaluation	53	47
	Average	39	61

Chart 4.3	Results for QP Area 2.3, appropriateness of the overall evaluation methods

The evaluation methods should be clearly outlined in the report, and their appropriateness, relative to the evaluation's primary purpose, focus and users, should be explained, pointing out the strengths and weaknesses of the methods.

■ Good	■ Satisfactory	Year review[16]	Unsatisfactory ■ Poor □
11%	48%	**2004**	37% 4%
5%		2003	95%
7%		2002	93%
22%		2001	78%
16%		2000	84%

◄ Satisfactory and Good   Unsatisfactory and Poor ▷

Chart 4.4	Results for QP Area 2.1, nature, make-up and appropriateness and biases of the evaluation team

a) The report should outline the nature (eg external or mixed) and make-up of the team (eg sectoral expertise, local knowledge, gender balance) and its appropriateness for the evaluation

b) The evaluation report should outline the evaluator(s)' biases that might have affected the evaluation and how these have been counteracted

■ Good	■ Satisfactory		Unsatisfactory ■	Poor □
2%	23%	**2004 (2.1a)**	60%	15%
2%	7% 12%	**2004 (2.1b)**		79%

Despite this improvement, it still remains rare for evaluators to include a full description of methodologies in their reports, particularly details of any survey tools that may have been used, and to discuss the strengths and weaknesses of their approaches. USAID's evaluation of its humanitarian response in the DRC during 2000–2004 provided a good description of the methodologies used in that country's difficult context (Box 4.4).

Evaluations of humanitarian action rarely use control groups, often because it is difficult in many emergency situations to identify appropriate groups who have not been touched in some way by the relief effort. An exception is the American Refugee Committee's evaluation of income-generation projects for Liberian refugees in Guinea, a relatively controlled situation with a well-defined target group of beneficiaries (Box 4.5, p 140).

Box 4.4	Good practice in describing evaluation methodology

Questionnaire for Management Informants A management questionnaire was developed and used as a guide for interviewing informants on management aspects. Approximately 60 management staff were interviewed, including staff from the US and DRC governments and from local and international NGOs. The questionnaire was sent by email to approximately 40 people who could not be personally interviewed. Ten people completed and returned the questionnaire and others were followed up with phone interviews. The responses to the written questionnaires were tallied and the results synthesized. The individual responses in both interview and written form are confidential.

Focus Group Interviews A "Community Questionnaire" was created to guide focus group discussions with program beneficiaries, including questions related to satisfaction with the assistance received, participation, impact and sustainability. The results were tallied and synthesized.

Interviews with Congolese NGOs A questionnaire was developed for the purpose of reaching staff of local NGOs, many of whom had partnered with CARE and IRC to implement OTI and OFDA funded programs. Ten NGOs were interviewed and the interviews were compiled and analyzed.

Sampling Technique Sampling was purposive. The intended design of the field visits had to be modified somewhat due to the logistical constraints described above. Interviews took place in 18 locations. Those interviewed covered a variety of beneficiary and other groups, including some that had not received assistance.

Source USAID, 2004, pp 2–3.

Evaluators

The importance of finding good evaluators has been highlighted in earlier ALNAP Reviews and remains an issue of concern for evaluation managers. Information about the nature and make-up of evaluation teams is poorly covered in most reports, as are declarations of known biases or possible conflicts of interest (Chart 4.4). It is often not possible to tell from the reports to what extent any of the team members had a good knowledge of the country or how any language problems were

overcome. There is also a question of transparency: readers should know whether evaluators have any potential conflicts of interest.

Sida's evaluation of assistance to IDPs in Indonesia is unusual in containing a clear description of the backgrounds and specialisations of the five team members involved in the evaluation. In the food security report of ECHO's evaluation of its Zimbabwe operations in 2002–2003, the evaluator declared his experience as an agricultural economist and his strongly held views on the technical advisability of growing maize in most parts of Zimbabwe. This helps the reader to understand the basis on which his recommendations about the future direction of the programme were made.

Using evaluators who have had an involvement with the programme at an earlier stage can create problems. In one evaluation in this year's set, consultants were used who had been involved in the original design of the programme. Perhaps surprisingly, their report turned out to be highly critical, so much so that staff members and key stakeholders virtually refused to consider the findings and recommendations.

Box 4.5	Experience of using a control group – the American Refugee Committee's evaluation of income generation projects for Liberian refugees in Guinea

A control-group survey was also conducted in Laine camp, among persons doing a business in the camp but who had not benefited from any loan or grants program, either of ARC or any other organization. The survey questions for the beneficiary-group and the control group are mainly the same… The beneficiary and the control group are thus both composed of persons who started/ or are doing a business, but who differ in whether they benefited from a grant or a loan. (ARC, 2004, pp 5–6)

The beneficiary survey results suggested a significant impact of the basic loan programme. In the control group, a portion of respondents achieved similar or even slightly better results than the beneficiary group. However, the lack of vulnerability indicators meant that there was no hard evidence that the control-group respondents were less vulnerable. The evaluation report therefore concluded that the findings on the impact of the programme were somewhat inconclusive.

However, the independence of the evaluation team from key stakeholders does not necessarily correlate with the usefulness and importance of the final report. A system-wide evaluation of the response to the 2002–2003 drought emergency in Ethiopia was run by a steering committee including representatives of government, UN agencies, donors and NGOs. The team included personnel who had been involved with key agencies but also included an independent consultant of high reputation. The preface to the report noted that 'The compilation of the report was part of the consultative process, which required compromise and, on occasion, difficult decisions' (DPPC/SDMT/UN, 2004, p 4). It does however seem that the team and the process produced a useful report that has and is influencing subsequent emergencies in Ethiopia by shifting the analysis of emergencies away from an almost exclusive focus on food supplies.[16]

Most evaluations in this set were carried out without the involvement of national consultants as part of the team. The orientation is still very much that of Northern-based consultants going out to assess humanitarian action in the South. Exceptionally, the joint-agency evaluation of provision of seeds to drought-affected people in eastern and central Tigray was carried out by national evaluators. The programme and the evaluation was a joint effort between the Italian NGO CISP, the Relief Society of Tigray, with funding from USAID–OFDA. This issue highlights commissioning agencies' policies and practices on the make-up of evaluation teams, as well as a possible lack of capacity building and mentoring of Southern evaluators.
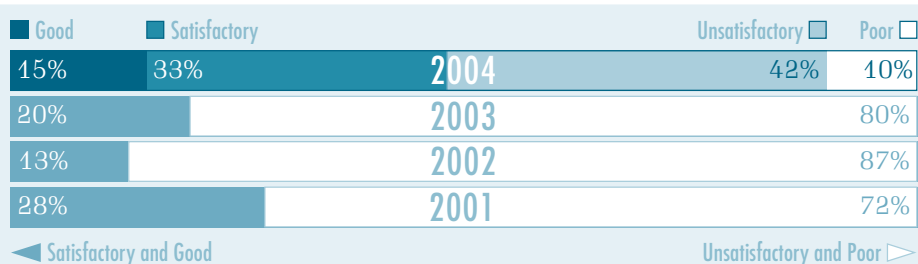
Involvement of primary stakeholders

Consultation with primary stakeholders during EHAs has improved considerably, although just over half of the reports are still rated as unsatisfactory or poor in this area (Chart 4.5).

The Sida evaluation of assistance to IDPs in Indonesia offers an example of an investigation that aimed to give its primary focus to the views and experiences of the IDPs themselves. The findings were based on structured qualitative interviews with IDPs and former IDPs, in combination with participant observation and analysis of documentary literature from the international aid community. The focus group interviews were differentiated by gender. An annex in the report gives full details of the context of each location where surveys were carried out (SIDA, 2004).

| Chart 4.5 | Results for QP Area 2.4a, consultation with and participation by primary stakeholders |

The evaluation report should outline the nature and scope of consultation with, and participation by, beneficiaries and non-beneficiaries within the affected population in the evaluation process. (A satisfactory or higher rating should be given only where evidence is presented of adequate consultation and participation of primary stakeholders in the evaluation process, or where, in the assessor's view, it has been successfully argued as inappropriate due to security or other reasons.)

■ Good ■ Satisfactory   Unsatisfactory ■   Poor □

15%	33%	**2004**	42%	10%
20%		**2003**		80%
13%		**2002**		87%
28%		**2001**		72%

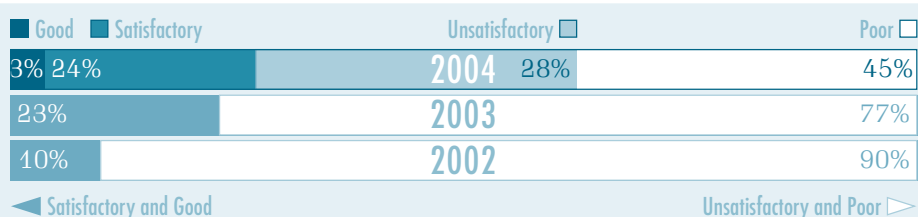◄ Satisfactory and Good   Unsatisfactory and Poor ▷

Use of international standards

The satisfactory use of international standards, such as the Sphere minimum standards, as methods of assessment in evaluation reports occurs in just over a quarter (27 per cent) of reports in the 2004 set. This is an area that has shown only limited improvement since 2002 (Chart 4.6). The poor performance in this area must be partly attributed to the weakness of terms of reference in stipulating the expectation of good evaluation practice, as noted in Section 4.6.1 above. This may in turn be linked to a generally weak application of standards by agencies to their operations on the ground.

| Chart 4.6 | Results for QP Area 2.5, adherence to international standards |

The evaluation report should assess the intervention against appropriate international standards (eg international humanitarian and human rights law, the Red Cross/NGO Code of Conduct, Sphere).

■ Good ■ Satisfactory   Unsatisfactory ■   Poor □

3%	24%	**2004**	28%	45%
23%		**2003**		77%
10%		**2002**		90%

◄ Satisfactory and Good   Unsatisfactory and Poor ▷

Box 4.6	Good practice in beneficiary consultation

Although time and conditions may not always allow for an ideal methodology, the WFP evaluation report on its Eritrea operations showed that a substantial amount of beneficiary consultation can be undertaken.

In the limited time available, the team visited 15 villages or camps, conducted in depth interviews with 60 households receiving relief rations, met with representatives from 33 Relief Committees (RCs), and visited 5 schools, 4 therapeutic feeding sites and 2 sites where supplementary feeding was taking place. The team employed a variety of methods to collect primary information including wealth ranking; … transect walks; seasonal calendars; … community mapping; MUAC [mid-upper arm circumference] measurements; group discussions with RCs, village leaders, kebabi administrators, school teachers, members of Parent Teacher Associations, pupils and health professionals; key informant interviews with local Eritrean Relief and Refugee Commission (ERREC) and IP staff; and informal interviews with 60 households benefiting from general feeding and 20 mothers of children receiving therapeutic or supplementary feeding. Villages and sites were chosen by the CO [country office] and SOs [sub-offices] to cover a variety of beneficiary types, agroecological zones and ethnic groups while balancing travel time with time available for consultation. Within any one community, households were purposively selected to explore the range of wealth and vulnerability. The findings of the team should in no way be considered statistically representative of households or beneficiaries. Household and school information was entered into simple databases to assist with analysis. (WFP, January 2004)

DEC evaluations maintain their general high standard, as noted in previous ALNAP Reviews. In the case of the Liberia appeal, the DEC evaluations continue to use the Red Cross/Red Crescent Code of Conduct, to which DEC agencies are signed up, as an overarching set of evaluation reference points, while, in this case, successfully weaving in the DAC criteria.

> ## Box 4.7   Using international standards
>
> The Danida evaluation of IDPs in Angola (Danida, March 2004) uses Sphere standards. Also, unusually, it looks at national standards, noting that the Angolan government is one of only two to have incorporated the Guiding Principles on Internal Displacement into national law, and goes beyond that to establish minimum standards for return. This evaluation has a detailed annex that compares Angolan standards in water, sanitation, health provision, non-food items and shelter with the equivalent in the Sphere minimum standards and the UNHCR and USAID field handbooks. This provides an overview of how these standards compare in relation to the programmes being reviewed

Constraints on evaluation

Just over 50 per cent of reports adequately describe the constraints that evaluators faced while carrying out their work (Table 4.2). Box 4.8, from a UNHCR-WHO joint evaluation gives details of some of the constraints that evaluators can face, and makes recommendations to the commissioning agency for the future.

Among the evaluation managers interviewed for this survey, evaluations were reported to have faced relatively few practical issues relating to the management of the work and the level of support given to the evaluators. The majority of evaluators found themselves well supported in the field, and security management was rarely a concern. Some managers noted that there was often a shortage of time, particularly lead and preparation time, although in discussion they doubted whether this had an adverse impact on the final quality of the evaluations in a significant number of cases.

The difficulty of finding evaluators, as mentioned above, may be linked with the importance of good evaluation planning that was mentioned by managers. As one put it, 'more time and effort before an evaluation pays back ten times. Every shortcut backfires'. Experience shows that at least three months may often be needed for the preparation of an evaluation before the evaluators start work, taking into account the time taken to agree the terms of reference and identify available evaluators.

About half of the evaluators interviewed expressed concern about the shortage of time available to do the job. In one case, the evaluator was given only 15 days in all,

Box 4.8	Practical constraints in evaluation

In the joint UNHCR-WHO evaluation of health and health programmes for Afghan refugees living in Pakistan, the report goes into some detail about the constraints faced by the team and as an example of good practice makes recommendations for future evaluations,

■ The team had very little influence over the evaluation schedule. In part, this was due to constraints of notification and security. This created obstacles at three levels: first, time was lost with initial formal plenary meetings and secondly, potential sampling biases were introduced. The choices of sites to visit were entirely pre-arranged, which, in conjunction with the high number of potential sites to visit and the already short duration of the evaluation resulted in potentially flawed sampling. Without accusing field staff of choosing showcase projects, there is no way to prove that this is not the case; random visits to un-scheduled sites were only possible occasionally. Thirdly, the fact that the visit of the team was announced beforehand everywhere created a certain 'police car effect'. The team complemented the schedule by visiting sites of concentration of urban refugees in the three cities visited.

■ Other methodological difficulties encountered were related to controls and 'representativeness' of key respondents. The problem of 'missing control groups' when investigating refugee populations is well described in the literature … and concerns the local host population and the non-displaced. A foray into Afghanistan was admittedly impossible to organize, and time constraints did not allow for visits to local settings. The evaluation had therefore to rely on previous personal knowledge (with regard to Afghanistan), local informants (with regard to Pakistan) and on available literature.

■ The problem of access to randomly chosen informants is well known, 'those who come forward or agree to be interviewed may not be representative at all'. In an extremely hierarchical society, it required considerable skills to get beyond the officials, i.e. views of the village elders or the most senior person within the group.

■ Although the team members were contacted well in advance about their potential participation in the evaluation, two of the consultants were given only one working day's notice before departure from home (the third team

CONTINUED

| Box 4.8 | Practical constraints in evaluation *continued* |

member received four working days' notice). The team composition, however, was extremely complementary. The only exception was the gender composition, which was a concern in a context where access to women respondents by male investigators is notoriously difficult, and where additional women on the team would have been desirable.

Conclusion

The only major methodological shortcoming of the evaluation was potentially flawed sampling; others, such as missing control groups, and difficulties of access to women as well as to genuinely 'representative' respondents, were to a great extent overcome.

Recommendations

■ Allow, considering the variety of both topics to cover and refugee settings to visit, for a slightly longer evaluation period in the field than two-and-half weeks.

■ Give earlier notice to the consultants than — as it was the case for two of them — one working day accorded between confirmation of contract and departure from home.

■ Tilt, (in a context where access to women respondents is notoriously difficult for male investigators), the gender balance of the evaluation team toward the female component.

■ Carry out such comprehensive evaluations of such costly health programmes earlier in their lifetime or rather, periodically.

■ Organize such comprehensive evaluations as 'staggered' evaluations. A previous or earlier field visit by the team leader could:

• Shape the schedule with a view towards more efficiency;
• Gather written data sources and write the narrative of the action;
• Improve sampling and design a methodology appropriate to specific
  constraints;
• Consider hiring a local consultant for technical topics; and
• Include scheduled writing-up time at the end of fieldwork for report.

UNHCR-WHO, December 2005, pp 20–21.

including all travel, the field visit and writing-up time, but agreed to do the work knowing that to guarantee quality he would have to donate a substantial amount of unpaid time. There is a sense that both staff and evaluators often have to work very hard to compensate for the lack of planned and allocated time. In the case of the WFP evaluation of its Afghanistan programme, the usefulness of a pre-visit by the team leader was noted both to facilitate planning and to generate interest and buy-in from staff and other key stakeholders.

4.6.3 Contextual analysis

Contextual analysis has continued to be a stronger area in two-thirds of reports, and this set offers some examples of good practice (Box 4.9). Some evaluation managers question whether all reports should contain a detailed context section, arguing that the immediate readers and users of the report will be familiar with the context. However, there are strong reasons for retaining an adequate context section:

• it shows the evaluators' analysis of the context in which the humanitarian response takes place;

• it gives the report a value and shelf-life to readers beyond the immediate circle of current stakeholders – to future staff members, archivists and researchers.

Box 4.9	Good practice in contextual analysis

WFP, Afghanistan

The WFP evaluation of its programme in Afghanistan looked at the context in which food aid was being distributed and the possible relationships to incentives to grow opium poppies. The report found that farmers do not make choices between wheat and opium production based on the relative price of the two crops.

> [F]armer decisions to grow poppy are driven by other considerations than food aid, such as: access to cash advances for inputs; relative profitability and risks of alternative crops; need to repay debts expressed in quantities of opium; fragmentation of holdings making poppy the only crop that

Box 4.9	Good practice in contextual analysis *continued*

produces enough income to feed a family from an average-sized farm of 2 jeribs (0.4 ha), and the fact that most of the wheat grown by small farmers is for own consumption with no surplus for sale and is therefore fairly unresponsive to variations in prices. (WFP, December 2004, p 56)

Terre des Hommes, Sri Lanka

The evaluators of Terre des Hommes's integrated Maternal and Child Health and Psycho-social programme in the Eastern Province of Sri Lanka (Terre des Hommes, June 2004) manage in eight pages to provide a comprehensive contextual analysis that includes:
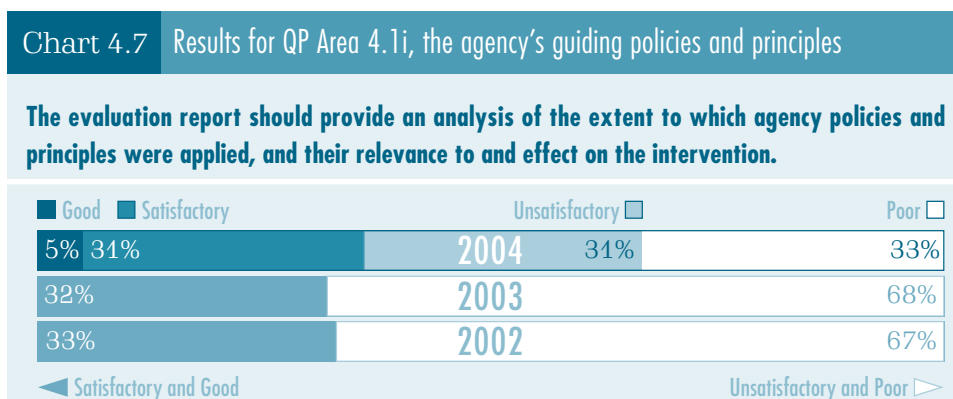
■ basic facts and humanitarian indicators for Sri Lanka;

■ political background to the conflict in Sri Lanka;

■ an overview of national health policies and strategies;

■ health indicators;

■ information on child rights, including child abuse and demobilised child soldiers;

■ regional context of the programme area in political, economic and health terms;

■ baseline health data.

Sida, Indonesia

The Sida evaluation of IDPs in Indonesia contains a detailed annex on the situation of IDPs, looking at the causes and impact of displacement as well as the rights of IDPs (Sida, August 2004).

UNHCR-WHO, Afghan refugees in Pakistan

The UNHCR-WHO evaluation of health and health programmes for Afghan refugees in Pakistan has an excellent context section that clearly places the refugees and the programmes targeting them in a historical, political and policy context (UNHCR-WHO, December 2003).

Chart 4.7 Results for QP Area 4.1i, the agency's guiding policies and principles

The evaluation report should provide an analysis of the extent to which agency policies and principles were applied, and their relevance to and effect on the intervention.

■ Good ■ Satisfactory	Unsatisfactory ☐	Poor ☐
5% 31%	2004 31%	33%
32%	2003	68%
33%	2002	67%

◄ Satisfactory and Good   Unsatisfactory and Poor ▷

4.6.4 Assessing the intervention

Institutional considerations

The application of agency policies remains a weak area, with two-thirds of reports not making adequate reference to the agency's humanitarian policies and guidelines that should govern the programme work (Chart 4.7). This weakness seems to stem in part from the failure of terms of reference to pinpoint this area as an important issue, and to list policies and guidelines that evaluators should bear in mind
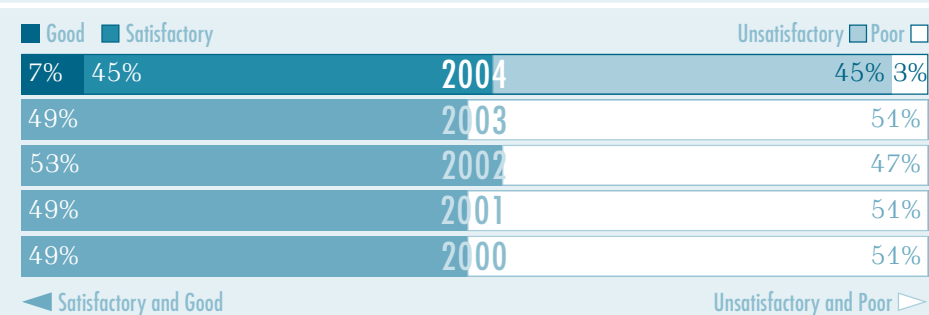
WFP's evaluation of its programme in Eritrea (WFP, January 2004) provides an example wherein organisational policies are considered, and programme performance measured against them. For example, in relation to WFP's commitment to women and enhanced commitment to women, the policy is used as a reference point in the text and in an annex. The WFP evaluation of emergency operations in Sudan (WFP, August 2004b) noted the lack of agency policy guidelines on recovery and transition issues.

Ratings over the years on agency management and human resources have remained steady, with about half of the reports being assessed as satisfactory in this area (Chart 4.8). Bearing in mind that many agencies place a high premium on getting the right people in the right place at the right time when responding to emergencies, it is surprising that human-resource and management issues are not featured more heavily in EHA reports. ECHO's institutional evaluation of Action Contre la Faim

looks at the NGO's human-resources policies and practices in some detail, with a dedicated annex covering the details of recruitment, training, turnover and remuneration.

Chart 4.8	Results for QP Area 4.1ii, the agency's management and human resources

The evaluation report should provide an analysis of the agency's management and human resource procedures and practices as applied, and their effect on the intervention. (This might include: level of experience/expertise of field staff; use of national and expatriate staff; staff turnover; field/HQ communications and relations; briefing and debriefing procedures; training and learning practices; security.)

■ Good ■ Satisfactory   Unsatisfactory ■ Poor □

Year	Satisfactory and Good		Unsatisfactory and Poor
2004	7%	45%	45% 3%
2003	49%		51%
2002	53%		47%
2001	49%		51%
2000	49%		51%

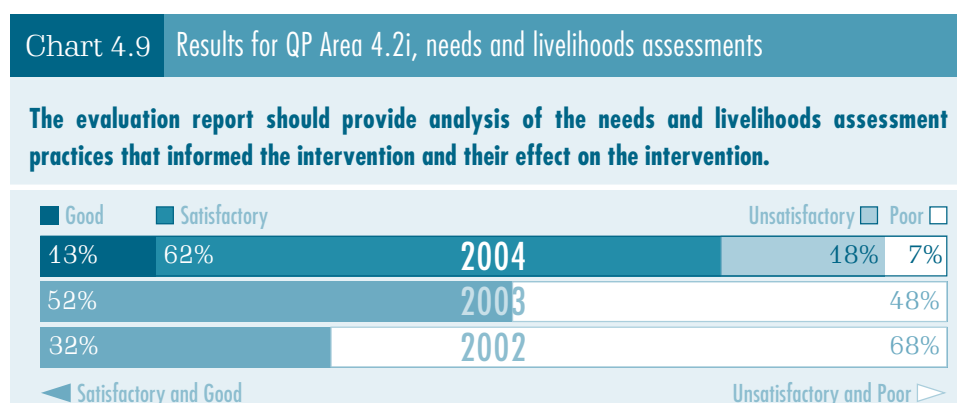◄ Satisfactory and Good   Unsatisfactory and Poor ▷

Needs assessment, objectives, planning and implementation

This element of the section on assessing the intervention, in the quality proforma, is one of the highest rated (Table 4.3).

Table 4.3	Results summary for QP Section 4.2, needs assessment, objectives, planning and implementation

		Satisfactory or good (%)	Unsatisfactory or poor (%)
4.2.i	The needs and livelihoods assessments that informed the intervention	75	25
4.2.ii	Intervention objectives	80	20
4.2.iii	Programme cycle processes – the evaluation report should analyse the following processes and their effect on the intervention:		
	a. planning	46	54
	b. implementation	78	22
	c. monitoring and/or real-time evaluative mechanisms	82	18
	d. intervention expenditure	45	55
	Average	68	32

The analysis of needs and livelihoods assessments in evaluation reports has improved substantially since 2003 (Chart 4.9). This positive trend is encouraging but is not reflected in agency performance in this area of needs assessment. Darcy and Hofmann note that good assessment practice is about having enough relevant information on which to base sound analysis and judgements (Darcy and Hofmann, 2003, p 2). Box 4.10 illustrates some of the problems in this area, with findings from two evaluations.

Chart 4.9 Results for QP Area 4.2i, needs and livelihoods assessments

The evaluation report should provide analysis of the needs and livelihoods assessment practices that informed the intervention and their effect on the intervention.

■ Good	■ Satisfactory		Unsatisfactory ☐	Poor ☐
13%	62%	**2004**	18%	7%
52%		**2003**	48%	
32%		**2002**	68%	

◄ Satisfactory and Good   Unsatisfactory and Poor ▷

Box 4.10 Problems with needs assessments

UNHCR, East Timor

Regarding the repatriation and reintegration programme of refugees in East Timor, UNHCR's evaluation found that there was a lack 'of any systematic needs assessments, whether by village, sub-district or district'. The focus on housing 'was based on the knowledge that some 70 per cent of housing had been destroyed, rather than a discussion with the vulnerable categories it was intended to target' (UNCHR, February 2004, 32).

Danida, Angola

Danida's evaluation of assistance to internally displaced persons in Angola (Danida, March 2004) found that agencies did not appreciate the increased importance of revised needs assessment in the post-conflict situation. For example, distribution programmes for seeds and tools were a continuation of basic relief responses undertaken during the war, without looking at supporting the establishment of commercial markets in seeds and tools.

Darcy and Hofmann also note that there is a wide range of approaches to the assessment of food security, and wide variation in the methodologies adopted by different agencies to collect data (Darcy and Hofmann, 2003, p 4). This is borne out by findings from some of this set of evaluation reports, summarised in Box 4.11.

Box 4.11	Needs assessment in contexts of food insecurity

DPPC/SDMT/UN, Ethiopia

The joint-agency evaluation of emergency response in Ethiopia (DPPC/SDMT/UN, October 2004) focused on the fact that the developing drought emergency was being analysed mainly in terms of crop production: 'There was an initial lack of appreciation that the crisis was far more complex than food availability. Assessment missions had focused on crop production rather than access by the affected populations to food, water, health, seeds and basic veterinary services.' The report recommended the introduction of livelihoods analysis into early warning systems. The report notes that past emergency responses in Ethiopia concentrated on food rations and, with this 'food first' bias, insufficient attention was given to health and nutrition.

WFP, Sudan

A WFP evaluation looked carefully at the needs assessment being used by WFP and other agencies in Sudan, particularly the shortcomings of the annual needs assessment exercise (WFP, August 2004). While being recognised by key stakeholders as valuable, the report recommended moving to a lighter, more iterative needs assessment process that would be less cumbersome, more sensitive and more credible.

A number of reasons have been put forward for the lack of adequate needs assessment by agencies, as follows (Darcy and Hofmann, 2003).

• The underlying issue of a lack of a shared definition of 'humanitarian need', and lack of universal benchmarks and standards to define the severity of a situation and the response requirements.

• Needs assessment, at least in the formal sense, often plays only a marginal role in the decision-making of agencies and donors.

- Monitoring is typically focused on the input–output equation of project management, rather than on an assessment of the external environment and the changing nature of risks.

- Overwhelmingly, needs assessments are conducted by operational agencies, often to substantiate funding requests. This allows for the close correlation of needs analysis with the design and execution of responses, but raises major questions about objectivity of analysis.

Darcy and Hofmann conclude that needs assessment as currently practised is inadequate to provide the information upon which to base genuinely impartial responses. Too little priority is given to the process of assessment throughout the course of a crisis; and it is too closely aligned to 'front-end' fundraising processes (Darcy and Hofmann, 2003, p 6).

Generally the evaluation reports in this set have been rated well in terms of their assessments of the relevance of intervention objectives (Table 4.3). This area is of crucial importance for evaluators as it is possible to assess effectiveness only when there is a clear statement of objectives. Good practice in this area can involve an analysis of the logical frameworks that are widely used by agencies to plan programme responses (Box 4.12).

| Box 4.12 | Programme objectives – analysing logical frameworks |

Tearfund, Sierra Leone (Tearfund, February 2004b)

The programme objective taken from the logical framework was: 'To enable resettlement, recovery and rebuilding of key communities in Jaluhan Chiefdom, which has been devastated following ten years of conflict'. The evaluation compared the programme with the Government of Sri Lanka's National Recovery Strategy and the UN Development Assistance Framework (UNDAF) for Sri Lanka.

However, the evaluation report noted that the word 'key' was not clearly defined, and potentially created ambiguity and impeded measurement of the achievement of this objective. Future formulations of objectives should contain more explicit criteria for prioritisation.

CONTINUED

| Box 4.12 | Programme objectives – analysing logical frameworks *continued* |

Danida, Angola (Danida, March 2004)

The evaluator here states that objectives can often best be stated in a logical framework format, but notes that these can be of poor quality. The report cites an example: 'the target population should achieve "a level of mortality below 1% among the general population and less than 2% for those under 5 years of age".' This is meaningless as there is no reference to the existing mortality rates, and no reference to the term over which the percentage is to be measured. The Sphere minimum standards state that when baseline data are not known, agencies should aim to maintain the crude mortality rate at below 1.0/10,000 per day and 2.0/10,000 per day for children under five years old. However, the report observes that, before 2003 in Angola, large changes in humanitarian conditions due to changing military situations meant that agencies were of necessity reactive rather than proactive, and very detailed logical frameworks would not have been useful.

Water and sanitation in Zimbabwe (ECHO, March 2004b)

Reviewing the logical framework, the evaluator here notes general confusion of project objectives, results and activities. A number of reports also note the lack of objectively verified indicators.

Liberia crisis appeal (DEC, August 2004)

The evaluator observes a surprising lack of logical frameworks for the DEC proposals, given that most if not all DEC member agencies use logical frameworks (or log frames) as planning tools. The report notes that log frames would have made reporting quicker for the agencies, and more accurate, and would have contributed to better monitoring.

The evaluation reports' analysis of programme-cycle processes is varied, as shown in Table 4.4. Implementation and monitoring score well, but less than 50 per cent of the planning and expenditure areas are rated as satisfactory or good. The reasons for the rating differentials shown in Table 4.4 are not entirely clear. It may be that both evaluators and evaluation managers understandably concentrate on the implementation phase of the programme cycle as being a core part of the intervention. The quality of monitoring and evaluation will also be of particular concern to evaluators.

Table 4.4	Results for QP Section 4.2iii, programme cycle processes

The evaluation report should provide analysis of the following processes and their effect on the intervention. (Consideration in this analysis should be given to local capacities; primary stakeholder consultation and participation; local and national partnerships.)

	Satisfactory or good (%)	Unsatisfactory or poor (%)
4.2.iii (a) Planning	46	54
4.2.iii (b) Implementation	78	22
4.2.iii (c) Monitoring and/or real-time evaluative mechanisms	82	18
4.2.iii (d) Intervention expenditure	45	55
Average	**63**	**37**

Neglecting to analyse the planning area of the programme cycle is unfortunate, as this feature, together with a sound needs assessment and well-defined and relevant objectives, is likely to be one of the key determinants that produce an effective and appropriate humanitarian response. The Terre des Hommes Sri Lanka report (TDH, June 2004) has a specific section on programme management, and notes that planning documents and the logical framework were not adapted to meet fast-changing circumstances, caused partly by a ceasefire in the project area.

DAC Humanitarian evaluation criteria

Chart 4.10 shows that the use of the DAC criteria continues to strengthen as a core component of EHA practice.[18] Table 4.5 shows the ratings for each of the DAC criteria for 2004. Despite this relatively positive overall rating of achievement for the DAC criteria, it should be noted that six of the thirty reports in this set did not explicitly use the DAC criteria, although in some cases the criteria were in fact covered. It is encouraging that the evaluative analysis of effectiveness has continued to maintain high ratings, with 85 per cent of reports dealing with this criterion in a satisfactory or good way (Chart 4.11).

As an example of good practice, the evaluation of USAID's humanitarian response in the Democratic Republic of Congo (DRC) for 2000–2004 (USAID, September 2004) has an extensive section on effectiveness. The section starts by defining effectiveness as the extent to which USAID's programmes achieved their purpose, and flagging up the issues to be covered:

Chart 4.10	Results for QP Area 4.3, DAC criteria overall[19]

The evaluation report should provide evidence of an adequate application of standard evaluation of humanitarian action criteria as per the OECD/DAC definitions.
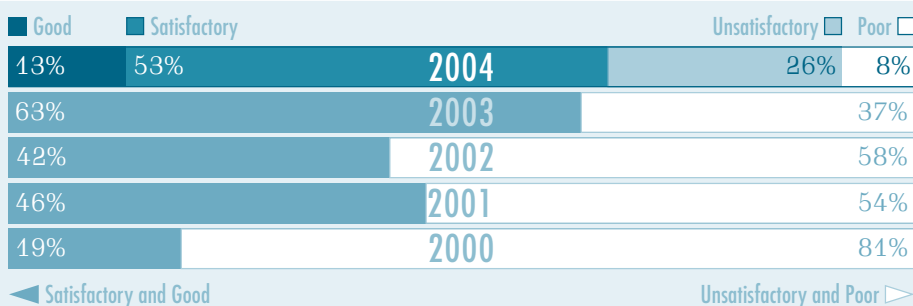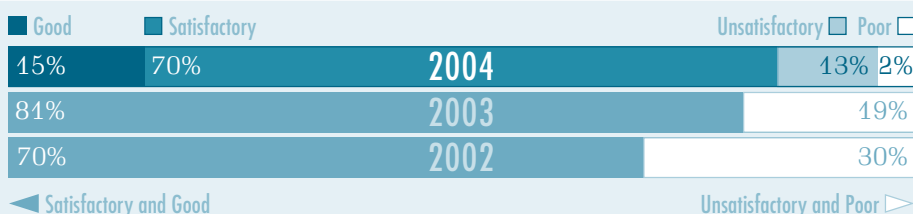
■ Good	■ Satisfactory		Unsatisfactory ■ Poor □
13%	53%	**2004**	26% 8%
63%		**2003**	37%
42%		**2002**	58%
46%		**2001**	54%
19%		**2000**	81%

◄ Satisfactory and Good　　　　　　　　　　　　　　　　　　Unsatisfactory and Poor ▷

Table 4.5	Results summary for QP Area 4.3, individual DAC criteria

		Satisfactory or good (%)	Unsatisfactory or poor (%)
i.	Efficiency	60	40
ii.	Effectiveness	85	15
iii.	Impact	40	60
iv.	Relevance/appropriateness	82	18
v.	Sustainability/connectedness	84	16
vi.	Coverage	50	50
vii.	Coherence	62	38
	Average	**66**	**34**

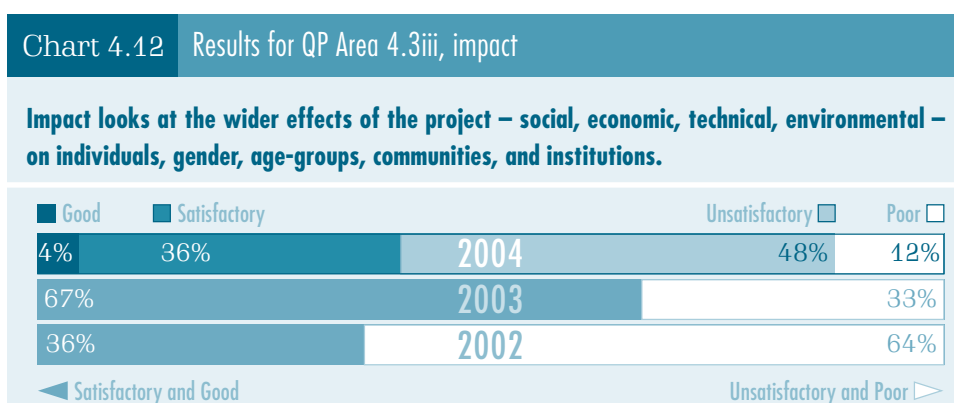| Chart 4.11	Results for QP Area 4.3ii, effectiveness

Effectiveness measures the extent to which the activity achieves its purpose, or whether this can be expected to happen on the basis of the outputs. Implicit within the criteria of effectiveness is timeliness of the response. (Although coordination is not a formal criterion, the OECD/DAC guidance suggests that given its importance, it should be considered under this criterion.)

■ Good	■ Satisfactory		Unsatisfactory ■ Poor □
15%	70%	**2004**	13% 2%
81%		**2003**	19%
70%		**2002**	30%

◄ Satisfactory and Good　　　　　　　　　　　　　　　　　　Unsatisfactory and Poor ▷

- USAID's DRC strategy for protection and advocacy;
- USAID's programme plans;
- timeliness of operations;
- struggling for access;
- uses of food aid;
- gender issues.

The DAC definitions of humanitarian evaluation criteria expect that coordination will be treated under effectiveness, although it is often treated under coherence, as is the case in the above-mentioned USAID report. Lack of effective coordination between agencies is often cited as a weakness of major humanitarian responses and has been the subject of many studies. It is therefore important that coordination is analysed in evaluation reports to test agencies' commitments to effective coordination. Box 4.13 gives examples of good practice for evaluation in covering coordination issues.

In contrast to effectiveness, the assessment of the impact criterion is now the weakest area within the DAC criteria, with only 40 per cent of reports achieving satisfactory or good ratings, including only one report receiving a 'good' rating (Chart 4.12). This is perhaps not surprising given the lack of baseline data, clear objectives and adequate monitoring and evaluation systems that evaluators often find, although it is not clear why the ratings for 2003 were much higher than for 2004 or 2002.

Chart 4.12	Results for QP Area 4.3iii, impact

Impact looks at the wider effects of the project – social, economic, technical, environmental – on individuals, gender, age-groups, communities, and institutions.

■ Good	■ Satisfactory		Unsatisfactory ■	Poor □
4%	36%	2004	48%	12%
67%		2003		33%
36%		2002		64%

◄ Satisfactory and Good   Unsatisfactory and Poor ▷

| Box 4.13 | Covering coordination |

CARE emergency programme, West Bank and Gaza

This report (CARE, May 2004) notes that agencies working in the West Bank and Gaza face special challenges in relation to effective coordination. These are summarised as:

■ general lack of capacity and organisation in the Palestinian Authority, with key officials sometimes being overloaded;
■ certain confusion and competition within UN agencies in the West Bank and Gaza;
■ ad hoc donor coordination mechanisms;
■ uncertainty about how international NGOs fit into the various coordination groupings.
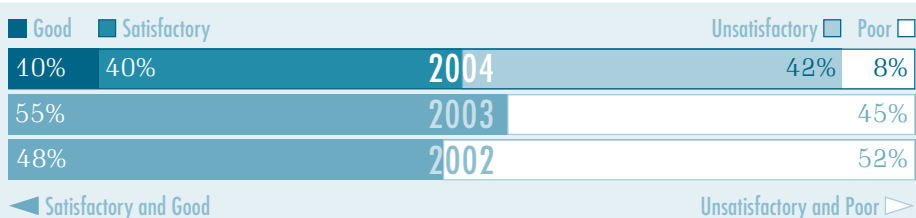
The report recommends, among other things, that CARE should map out and prioritise its commitments to and resourcing for coordination, and rationalise and clarify its wide range of partner relationships.

UNHCR-WHO: Afghan refugees in Pakistan

This evaluation report (UNHCR-WHO, December 2003) goes into some detail about coordination within the health sector for Afghan refugees living in Pakistan, noting that lack of coordination can lead to duplication, re-inventing the wheel and a lack of standard protocols for clinical practice and training. It concludes that regular, structured coordination meetings need to be introduced.

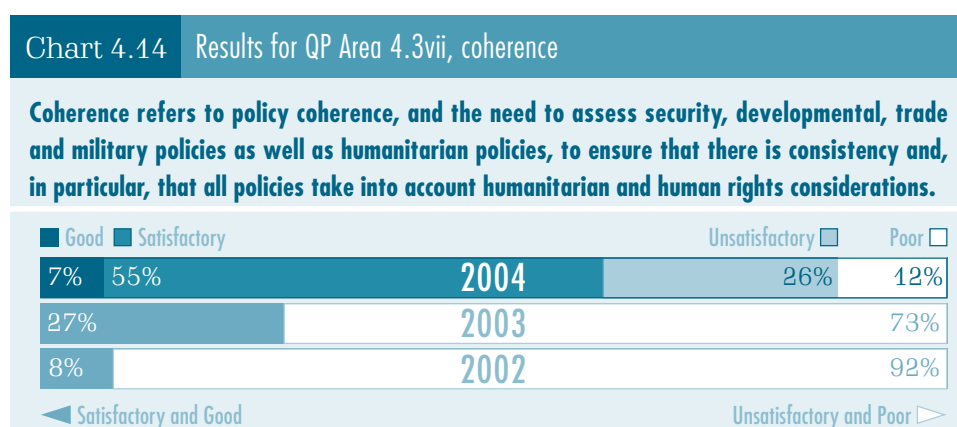| Chart 4.13 | Results for QP Area 4.3vi, coverage |

The need to reach major population groups facing life-threatening suffering wherever they are, providing them with assistance and protection proportionate to their need and devoid of extraneous political agendas.

■ Good ■ Satisfactory   Unsatisfactory ■ Poor □

	Good	Satisfactory		Unsatisfactory	Poor
2004	10%	40%		42%	8%
2003	55%			45%	
2002	48%			52%	

◄ Satisfactory and Good   Unsatisfactory and Poor ▷

The evaluation of the American Refugee Committee's income generation programme for Liberian refugees in Guinea moved beyond outcomes and economic impact and, by survey, assesses the social impact of the scheme: 'Respondents mention greater social status, not having to borrow money anymore, being healthier and having more pride, more often than having more money' (ARC, June 2004, p 17). As a result of the scheme, 20 per cent of respondents mentioned that there was less conflict in their families.

Coverage by interventions is another area that is rated poorly for about half of the evaluation reports (Chart 4.13). Part of the reason for this result may be that some agencies, particularly those with both development and humanitarian interests, use the standard DAC development criteria that do not include coverage, the latter being specifically added to the DAC guidelines on humanitarian evaluation (OECD/DAC, 1999). Box 4.14 (overleaf) gives two examples of coverage being addressed appropriately in evaluation reports.

The generally positive trend in the treatment of the DAC criteria is illustrated in areas such as coherence, which has traditionally been poorly covered (Chart 4.14). In some reports there continues to be a lack of clarity about the definition of coherence. It is often confused with coordination, which in the DAC definitions is dealt with under effectiveness.

Chart 4.14	Results for QP Area 4.3vii, coherence

Coherence refers to policy coherence, and the need to assess security, developmental, trade and military policies as well as humanitarian policies, to ensure that there is consistency and, in particular, that all policies take into account humanitarian and human rights considerations.

■ Good ■ Satisfactory		Unsatisfactory ■	Poor ☐
7% 55%	**2004**	26%	12%
27%	2003		73%
8%	2002		92%

◄ Satisfactory and Good   Unsatisfactory and Poor ▷

The evaluation of the DEC agencies' response in the Liberia crisis of 2003/04 (DEC, August 2004) looked at coherence partly in terms of sustainability and vulnerability reduction. The report also notes that several NGOs were strongly involved in influencing the planning and practice of refugee reintegration by the government and UN agencies: 'Some DEC members have strong links with pressure groups seeking to change laws regarding SGBV [sex and gender-based violence] and facilitate meetings of Liberian human rights, pro-democracy and peace-building NGOs to develop and implement strategies for influencing work on transitional justice in Liberia' (DEC, August 2004, p 28).

Box 4.14   The issue of coverage

Coverage is highlighted as a crucial issue in the DEC evaluation of response to the Liberia crisis (DEC, August 2004). First, the report notes that the UN appeals were extremely poorly funded. Second, security problems severely restricted agencies' access to populations in need:

> The lack of access to the IDPs behind rebel lines raised concerns regarding their condition and well-being. Fortunately the worst fears did not materialise, or at least not to the level of a humanitarian disaster that may have been expected. This was due to the often long-standing survival strategies of Liberians living in the more remote counties, such as crossing the border or hiding in the forests and bush, and the high availability of wild food (bush meat, fruits etc) in the fertile rural areas of Liberia. (DEC, August 2004, p21)

The evaluation notes the problem of coverage for intensive work such as psychosocial and DDRR [Disarmament, Demobilisation, Rehabilitation and Reintegration] projects. Probably only a tiny proportion of the women and children who were victims of extreme abuse were reached.

> Security and access, the main constraints to adequate coverage, dominated the agencies' activities well into the programme period. What is of note, is the variation in security analysis by the different agencies that produced a related variation in access and coverage … this raises questions regarding shared analysis and a common attitude to risk, both for staff and beneficiaries. (DEC, August 2004, p22)

| Box 4.15 | Covering protection from gender-based exploitation and violence |

The UN Secretary-General's *Bulletin on Special Measures for Protection from Sexual Exploitation and Sexual Abuse*[20] and the IASC Statement of Commitment on Action to Address Gender Based Violence in Emergencies have further focused the attention of agencies on protection issues following the UNHCR/Save the Children report in 2002 (UNHCR and SCF, 2002) on sexual violence and exploitation of refugee children in West Africa.

Only the evaluation reports dealing with programmes in West Africa have looked at protection issues in the light of the UN and inter-agency commitments referred to above. The DEC Liberia report (DEC, August 2004) notes that several DEC agencies included variations of the IASC code on sexual exploitation in their staff contracts, prohibiting sexual contact with beneficiaries, and recommends that all DEC agencies should incorporate elements of the IASC code prohibiting the sexual exploitation of aid recipients in their staff contracts. Agencies working with partners should incorporate such elements in their formal contractual agreements and memoranda of understanding. The report also notes the importance of conceptualising protection activities so that effective strategies can be developed to include strong advocacy elements.

The same evaluation report considers that probably only a tiny proportion of the women and children who were victims of extreme abuse could be contacted, and comments:

> As the stigma of rape is so considerable, there is a risk that the women and children would not attend for fear of exposure to potentially disastrous repercussions such as rejection by their families and communities. The infiltration of the camps by ex-combatants (i.e. the perpetrators in most cases), could even place the women at risk of reprisal should those men suspect [them] of seeking legal redress. (DEC, August 2004)

Chart 4.15 — Results for QP Area 4.4i, the use of international standards

a) **The evaluation report should assess the extent to which relevant international standards were used in the planning, implementation and monitoring of the intervention** (eg international humanitarian and human rights law; the Red Cross/NGO Code of Conduct and developing standards, eg Sphere).

b) **The evaluation report should assess the intervention against appropriate international standards** (eg international humanitarian and human rights law; the Red Cross/NGO Code of Conduct, Sphere).
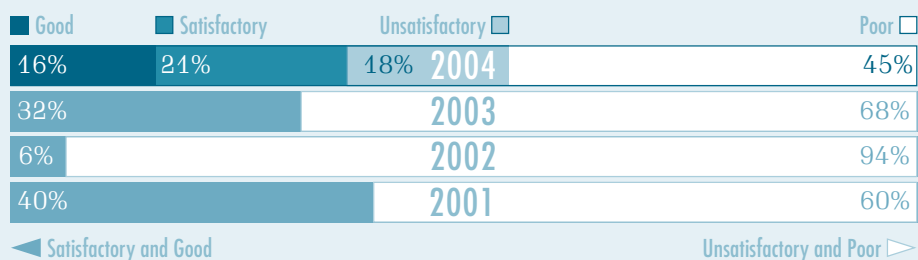
■ Good	■ Satisfactory	Unsatisfactory ■		Poor □
3%	15%	15%		67%
14%			2003	86%
17%			2002	83%

◄ Satisfactory and Good   Unsatisfactory and Poor ▷
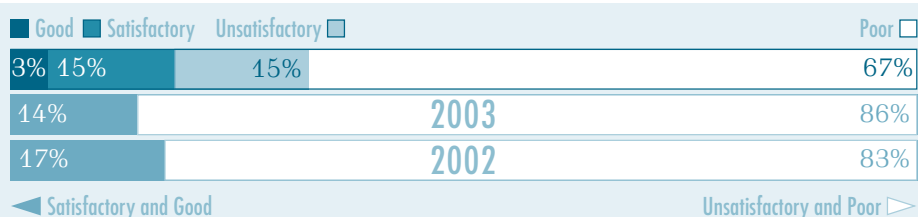
Table 4.6 — Results summary for QP Area 4.4, cross-cutting issues

		Satisfactory or good (%)	Unsatisfactory or poor (%)
4.4.i	International standards	18	82
4.4.ii	Gender equality	59	41
4.4.iii	Protection	38	62
4.4.iv	Capacity building	60	40
4.4.v	Advocacy	35	65
4.4.vi	Vulnerable/marginalised groups	45	55

Chart 4.16 — Results for QP Area 4.4iii, protection

The evaluation report should analyse the consideration given to protection throughout the intervention cycle, and the effect on the intervention.

■ Good	■ Satisfactory	Unsatisfactory ■		Poor □
16%	21%	18%	2004	45%
32%			2003	68%
6%			2002	94%
40%			2001	60%

◄ Satisfactory and Good   Unsatisfactory and Poor ▷

Cross-cutting issues

Cross-cutting issues remain one of the weakest sections of the EHA reports assessed. Although, encouragingly, gender equality and capacity building are satisfactory in about 60 per cent of reports, less progress has been made in other areas (Table 4.6). The lowest-rated criterion in this section relates to the assessment of the extent to which international standards were used in the planning, implementation and monitoring of the intervention (Table 4.6). This area has not improved significantly since 2002. (See also Box 4.7 and Chart 4.6 above, on the use of international standards.)

Progress is slow in improving coverage of protection issues (Chart 4.16), possibly surprisingly given the increased attention this topic is receiving within the humanitarian field. A number of reports make reference to the ALNAP guidance booklet on protection issues (Slim and Eguren, 2004).

The UNHCR evaluation of repatriation and reintegration in East Timor looks in detail at protection issues, noting the crucial importance of linking protection and programming activities, not seeing them as separate activities. The report found that 'a lack of systematic protection monitoring allows political and funding considerations rather than a consideration of objective conditions of returnees to dominate decision making, with negative repercussions for the best interests of refugees and returnees' (UNHCR, February 2004, p 97).

An evaluation of an MSF psychosocial project in Sri Lanka analysed a number of ways in which staff tackled protection issues:

- drawing the attention of the mandated UN agency to the issue of recruitment of child soldiers;
- finding ways of publicising human rights violations by non-state actors;
- pressing for stronger official action on child abuse (MSF-H, December 2003).

Advocacy is also often neglected in evaluation reports (Table 4.6), perhaps reflecting a situation in which advocacy can be poorly defined and recorded, although agency staff may give it high priority. Some evaluation managers felt that not all evaluations should deal with advocacy. However, there is also a view that advocacy initiatives always need to be considered as part of any humanitarian action, even if the resulting actions are low-key and localised, or in some cases not considered appropriate at all.

> ### Box 4.16 — Looking at advocacy work in the West Bank and Gaza
>
> The strategic review undertaken by CARE of its West Bank and Gaza programme (CARE, May 2004) is unusual within the set of evaluations for devoting a substantial section to advocacy. Although not specifically mentioned in the terms of reference, advocacy issues feature in the report because of the priority given to them by informants, and existing CARE policies and strategies, as well as because of advocacy activities already undertaken. The report notes that a CARE nutritional survey may have been responsible for changing donor priorities towards nutritional support to Gaza. The report contains extensive recommendations on the way forward for CARE West Bank and Gaza in a highly sensitive context, including that:
>
> - advocacy needs to be planned and strategised;
> - advocacy must be grounded in the fieldwork;
> - advocacy is a means to an end, and should not be confused with visibility;
> - advocacy strategies require careful context analysis, mapping and risk assessment.

Attention to vulnerable and marginalised groups remains steady, with just under half of the reports covering it satisfactorily (Table 4.6). Agencies and evaluators appear reluctant to look more broadly at the communities in which humanitarian action is taking place to see whether assistance is being provided equitably.

Darcy and Hofmann observe that:

> The notion of the 'vulnerable group' – typically based on assumptions about socio-economic status – can introduce artificial distinctions. Agencies and donors may concentrate resources heavily on a particular group while neglecting others. Not belonging to a 'vulnerable group' can itself be a major vulnerability factor. Assumptions about the needs and risks faced by particular groups may indeed be well-founded and based on previous evidence, but they should also be made explicit, and should be tested. (Darcy and Hofmann, 2003, pp 3–4)

The evaluation of USAID's humanitarian response in DRC raises this issue in relation to IDPs, noting that many agency personnel interviewed were opposed to developing a special focus on IDPs. They felt that this may result in inequity in assistance to vulnerable groups, pointing out that IDPs in DRC are often integrated into local communities in remote, inaccessible and insecure areas. The report details attempts to define and identify IDPs in terms of their: time of displacement, nature or reason for displacement, locations and shelter types, and types of assistance received. The report concludes that, in the DRC context, separate programmes for IDPs can result in unfairness that can strain relationships between IDPs and host communities.

> Borton et al (2005), reporting on learning from evaluations concerning support to IDPs, note that there is a 'strong vein of objection, not only to the ***treatment*** of IDPs as a separate category but even to their separate ***identification*** amongst all … vulnerable groups'. They recommend that this debate about the categorisation of IDPs be properly aired and resolved within the humanitarian sector, and that 'displacement be used as an indicator of potential vulnerability rather than as a means of defining target groups (Borton et al, 2005, pp 14–15).

Table 4.7 Results summary for QP Area 5.1, evaluation findings, conclusions and recommendations	Satisfactory or good (%)	Unsatisfactory or poor (%)
i. Use of secondary sources	43	57
ii. Conclusions	60	40
iii. Recommendations:		
a. Clear, relevant, implementable	80	20
b. Follow on from conclusions, and reflect consultation with key stakeholders	82	18
c. prioritised, with a timeframe for implementation, and suggested responsibility for follow-up	15	85

Box 4.17	Good practice in formulating recommendations

The evaluation of USAID's humanitarian response in the Democratic Republic of Congo, 2000–2004 (USAID, September 2004) is a wide-ranging study seeking to find more effective ways of targeting humanitarian assistance in the DRC and to better understand field practice in relation to the protection of IDPs and other vulnerable groups. The report underpins the IDP policy being developed by USAID and supports the joint-agency IDP evaluation initiative (see Section 4.4 above).

The report succeeds in coming up with just ten major recommendations that were considered practicable by the evaluators and within the capacity of the US government. Many of the recommendations were pitched at a policy level, and included the following:

- a local peace-making component should be included as part of USAID's humanitarian and 'preventive' strategy;
- USAID should focus on people most in need, rather than specifically IDPs;
- USAID should look in more detail at food-aid targeting, and strengthen its in-country analytical capacity, including increased use of Congolese staff;
- better linking between short-term humanitarian strategies and longer-term development goals;
- increased transparency on the part of USAID with regard to its strategies and policies for humanitarian assistance.

Tearfund's experience of earlier evaluations including up to 80 often-unfocused recommendations led it to press the evaluator of an IDP programme in Khartoum State, Sudan, for a focused and prioritised set of recommendations. The evaluation focused on the handover of the programme from Tearfund's implementing partner, the Fellowship for African Relief (FAR), to a national NGO, Sub-Saharan International Development Organisation (SIDO), and on capacity support work with that NGO. The report provides a good example of recommendations that are:

- clustered according to organisational responsibility, including the donor agency (Tearfund), the NGO intermediary (FAR) and the national NGO (SIDO);
- arranged in order of urgency, starting with issues to be addressed immediately;
- limited in number to 17 recommendations spread across the three agencies.

These focused recommendations are most likely to be of practical use to the managers and staff members responsible for implementing them.
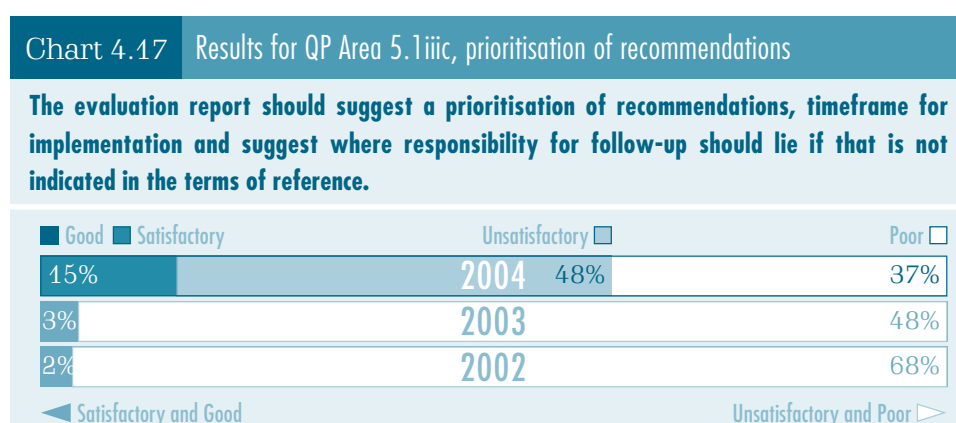
4.6.5 Assessing the report

Evaluation findings, conclusions and recommendations

In the majority of cases, managers felt that the conclusions of the evaluation reports were agreed without too much difficulty, although in some instances a lot of work had to be done to make the reports readable. As a result of past experiences, ECHO plans in future to ask the consultancy companies it contracts to use a report editor to overcome this problem.

In about half of the cases looked at, managers were satisfied that recommendations were presented in a way that facilitated their take-up and implementation. The quality proforma ratings in Table 4.7 show a more positive picture. However, the prioritisation of recommendations and their usability does remain an area of concern, with managers wary of reports with large unprioritised lists of recommendations. One noted that it was not uncommon for 50 per cent of recommendations to be of no use. Chart 4.17 reflects the very poor presentation of recommendations in terms of prioritisation, timetabling and allocation of responsibilities.

Dissemination and follow-up

In nearly all cases there was a process of report dissemination, even if that process was limited to a workshop in the agency's headquarters. After that, the Internet was the most common form of dissemination.

Chart 4.17	Results for QP Area 5.1iiic, prioritisation of recommendations

The evaluation report should suggest a prioritisation of recommendations, timeframe for implementation and suggest where responsibility for follow-up should lie if that is not indicated in the terms of reference.

Good ■ Satisfactory ■	Unsatisfactory ☐	Poor ☐
15%	2004 48%	37%
3%	2003	48%
2%	2002	68%

◀ Satisfactory and Good   Unsatisfactory and Poor ▷

| Box 4.18 | Getting the most out of the report |

In 2004 the American Refugee Committee (ARC) decided to evaluate a long-standing income- generation programme with Liberian refugees living in camps in the Forest Region of Guinea. The evaluation found that the programme was well thought-out and benefited from well-designed programme tools and good overall planning.

The ARC wanted to use the evaluation for advocacy purposes, to persuade donors that income generation for refugees was important. The report was widely circulated, put on the web and the evaluators and commissioning manager jointly wrote a case-study article for ODI's *Humanitarian Exchange*, promoting micro-finance initiatives for refugees (De Klerk and Nourse, 2004).

In most cases, funds were not specifically allocated within the evaluation budget for follow-up. It was also noted that terms of reference did not often spell out follow-up mechanisms and management responsibilities. Most managers reported that their agencies had, or were developing, a formal system for reporting on the follow-up of evaluation recommendations, with a designated management responsibility. This reflects an increasing focus on the usability of evaluation reports. Box 4.19 summarises some of the findings of a WFP study looking at the follow-up to evaluation recommendations.

However, it is not yet possible to give an accurate general picture of the extent to which evaluation recommendations in the reports covered by interviews have actually been followed up and implemented. This is partly because not enough time has elapsed and partly because managers did not have the information to hand, despite the follow-up systems being developed. The weakness of follow-up matrices is that, while they may show what post-evaluation activities have been undertaken, they do not show what impact the implementation has had, if any.[21]

The overall, anecdotal picture about the impact of evaluations is mixed. Not withstanding the positive findings from the WFP report described in Box 4.19, a minority of evaluation managers interviewed thought that the evaluations had an impact on programme response, human resource practices or partner development.

| Box 4.19 | Follow-up to evaluation recommendations |

In 2004 WFP undertook a study to look at the follow-up to evaluation recommendations commissioned between 2000 and 2002 in order to see to what extent recommendations were implemented (WFP, 2005). The study covered the full range of WFP's activities, not just its humanitarian work, and looked at recommendations from a sample of 26 evaluation reports. In the methodology for the study, learning was considered to have taken place when the recommendations of the evaluation were accepted and implemented by the management unit responsible. In cases where there was a follow-up project, the documents were examined to determine whether the recommendations had influenced the design of the new project.

The study found that the evaluations had made a significant contribution to WFP's knowledge base by generating acceptable, practical recommendations for project and policy improvement. Of all the recommendations, 95 per cent were reported as accepted by the management units concerned; 54 per cent had been implemented; and implementation was in progress for a further 34 per cent. Of all the accepted recommendations, 65 per cent were reflected in successor project documents. However, the report notes that in some cases implementation started only when the management unit responsible had been reminded of the recommendation by the study questionnaire.

WFP staff interviewed for the study thought that evaluation recommendations were influential and did improve project performance. However, the report notes that these judgements were intuitive and the results must therefore be treated with caution. The report recommended that WFP should institute a systematic follow-up procedure including a mechanism to report on the implementation of evaluation recommendations.

Managers reported only a few examples of the negative impact of evaluations, usually related to dynamics concerning the personality or approach of the evaluators. In most cases, managers were confident that the findings from evaluations had been fed into their agencies' knowledge management systems. Overall, therefore, managers were fairly satisfied with the outcomes of the evaluation processes that were discussed.

4.7 Conclusions

Evaluation managers agreed that three factors were overwhelmingly important for producing good evaluations:

1 good, realistic terms of reference with clearly stated objectives, adequate budget and realistic timeframe;

2 a good evaluation team;

3 ownership by the agency and stakeholders, including senior management, desks and country offices, and a commitment to learning.

Evaluators also stressed the importance of clear terms of reference and buy-in from the commissioning agency, its staff and partners, but were perhaps too modest to stress the value of good-quality evaluators.

This meta-evaluation shows that, although the evaluation of humanitarian action (EHA) has made progress in some areas, the overall quality leaves a great deal of scope for improvement. Evaluations have been strong in the areas of the use of the DAC criteria, and describing the context of the intervention; and there has been some improvement in the description and appropriateness of evaluation methods.

Weaknesses remain in the following areas of the evaluations assessed:

- terms of reference;
- descriptions of evaluation teams;
- consideration of agency policies;
- cross-cutting issues;
- primary stakeholder consultation (although this seems to be improving).

This meta-evaluation identifies the following trends in EHA:

1 increased efforts to get staff and senior management buy in to evaluations;

2   continued decentralisation of evaluation functions in some agencies to desk and country levels, getting evaluations seen as an essential part of project-cycle management;

3   more focus at headquarters level on thematic, sectoral and policy evaluations;

4   more emphasis on utilisation-focused evaluations and on the outcomes of evaluations;

5   a move to joint and inter-agency evaluations, as demonstrated by the UNHCR-WHO evaluations, the global IDP evaluation and the forthcoming Tsunami Evaluation Coalition thematic evaluations.

4.8 Issues for future ALNAP meta-evaluations

It is clear from this meta-evaluation, and from previous ALNAP meta-evaluations, that the development of the quality proforma and the annual review exercise have stimulated improvements in humanitarian evaluation processes. However, the following key questions should be addressed.

Is the quality proforma too demanding and detailed?

There is a feeling among some managers and evaluators that the quality proforma is too demanding and can only be fulfilled in the context of an expensive evaluation. There is mixed evidence on this issue from this year's group of evaluations. Some high-quality reports were produced by single evaluators. It seems likely that there has been some misunderstanding about whether every humanitarian evaluation should cover all the questions in the quality proforma. The proforma guidance notes often indicate that a criterion need not be met if a reason for exclusion is given. In practice, this option has very rarely been used.

Does the process of meta-evaluation focus too much on report quality?

Several respondents questioned the emphasis of the meta-evaluation process, wondering if too much attention was given to the quality of evaluation reports rather than outcomes and impact on humanitarian response.

> I have come around to the belief that what matters not so much is the professional quality of the evaluation, but the extent to which it influences agency policy. (evaluation manager)

> ALNAP is pushing too much towards quasi-research ... Evaluations should be instruments of organisational change. The evaluation industry is in danger of becoming self-serving. (evaluation manager)

This critique was the spur for the introduction into the meta-evaluation process of interviews with evaluation managers and evaluators. This year's initiative to consult with evaluators has been welcomed and should be widened next year, perhaps using some small, location-specific focus groups. However, there is nevertheless a strong argument for more systematic and thorough assessment of evaluation outcomes and impact.

Does the quality proforma cope well with thematic, organisational and policy evaluations?

Earlier attempts to make the quality proforma more flexible, in order to cover a wider variety of humanitarian evaluations, have not been successful, partly because of the wish to maintain the continuity of the datasets and continue to show trends over time.

Can the meta-evaluation cover a wider range of evaluative material?

Much of the evaluative material submitted for the meta-evaluation is Northern- and headquarters-focused. National organisations (governments and NGOs) and Southern evaluators are rarely included. ALNAP should consider light and cost-effective ways of extending its networking potential, perhaps starting with links to a regional centre like Nairobi.

Notes

1 Evaluation reports are submitted to the ALNAP Evaluative Reports Database (ERD) on a regular basis by agencies. The ERD is a fully searchable online database, which holds records of over 600 reports.

2 The author would like to acknowledge the contribution of peer reviewers Christian Bugnion, Peter Giesen, Simon Lawry-White and John Telford, as well as guidance from John Lakeman at the ALNAP Secretariat.

3 In 2001 the QP was made up of 37 questions in 11 sections. By 2004 it had grown to 50 questions consolidated in 8 sections. The 2005 version consists of 48 questions in 5 sections.

4 This revision process was carried out by the author, and peer reviewed by Chris Bugnion, Peter Giesen, John Telford and Simon Lawry-White

5 All the reports in this set were in English, except for one in French, partly reflecting limited language capacity within the assessment team.

6 Agreement means that the assessments were not more than one rating different, and did not cross the boundary between satisfactory and unsatisfactory.

7 A joint-agency evaluation in Ethiopia, including the government, UN agencies and international NGOs.

8 It should be noted at that four of the five ECHO reports make up a single country evaluation, on Zimbabwe, being three sectoral evaluations and a synthesis report.

9 Geographically, the reports covered Angola (x2), Coastal West Africa, DRC, Eritrea, Ethiopia (x2), Guinea (refugees), Liberia, Rwanda, Sierra Leone, Southern Africa, Sudan (x2), Zambia, Zimbabwe (x4), Afghanistan (x2), East Timor, Indonesia, Nepal, Pakistan, Sri Lanka (x3), West Bank and Gaza, with one global report (ACF's activities funded by ECHO).

10 The deadline for this evaluation set of the end of December 2004 meant that no evaluative work resulting from the Asian tsunami has been included in this Review.

11 Author's interview with UNHCR staff, February 2005.

12 Neither all of the IDP evaluations nor the synthesis report were available or were submitted for this year's meta-evaluation.

13 Formerly the DAC Working Party on Aid Evaluation.

14 The DAC Network has been reviewing the experiences and lessons learned from joint evaluations, and is due to publish a report, 'Joint Evaluations: Recent Experiences, Lessons Learned, Options for the Future' during 2005 (Breier, 2005).

15 Please note that all the figures in this section have been rounded to whole percentages.

16 Over the last five years, ALNAP has carried out 5 meta-evaluations. The meta-evaluations of 2000, 2001 and 2002 were published in the ALNAP Annual Reviews 2001, 2002 and 2003. Note that in 2004, the ALNAP Annual Review was renamed the ALNAP *Review of Humanitarian Action in 2003*. For the sake of clarity, the meta-evaluation results in this chapter are therefore referred to as follows: 2000 (*Annual Review 2001*); 2001 (*Annual Review 2002*); 2002 (*Annual Review 2003*); 2003 (*Review of Humanitarian Action in 2003*); and 2004 (*Review of Humanitarian Action in 2004*).

17 Personal findings of the author, during a visit to Ethiopia in April 2005.

18 It should be noted that one criterion for the selection of reports from the ALNAP database is that they use the DAC criteria.

19 Average ratings were taken for all seven elements of this measure (4.3.i–4.3.vii).

20 United Nations Secretariat, Secretary-General's Bulletin *Special measures for protection from sexual exploitation and sexual abuse* (ST/SGB/2003/13) http://ochaonline.un.org/GetBin.asp?DocID=1083

21 Danida is commissioning a systematic assessment of whether, in practice, evaluation recommendations have the desired effect on projects, programmes and policies. This work will focus on four or five countries, and a pilot study is due to be conducted in 2005.

References

Note The works listed here are sources additional to the 30 reports included in the meta-evaluation. Those 30 reports, referred to in the text by month and year of publication, are listed in full in Annex 3.

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

Borton, J, M Buchanan-Smith and R Otto (2005) *Support to Internally Displaced Persons – Learning from Evaluations: Synthesis Report of a Joint Evaluation Programme – Summary Version.* Stockholm: Sida.

Breier, H (2005) *Joint Evaluations: Recent Experiences, Lessons Learned, Options for the Future.* Presentation at the third meeting of the DAC Network on Development Evaluation, 2–3 June 2005.

Dabelstein, N (2003) *Note on Joint Evaluations.* Paper for the OECD/DAC Working Party on Aid Evaluation, 27–28 March 2003.

Danida (2003) *Framework for a Common Approach to Evaluating Assistance to IDPs.* Channel Research. Unpublished.

Darcy, J and C-A Hofmann (2003) *Humanitarian Needs Assessment and Decision-making.* Overseas Development Institute Humanitarian Policy Group Briefing 13, September. London: ODI.

Feinstein, O and GK Ingram (2003) *Lessons Learned from World Bank Experiences in Joint Evaluation.* Paper for the OECD/DAC Working Party on Aid Evaluation, 27–28 March 2003.

Griekspoor, A and K Burns (2003) Concept Paper *Establishing Inter-Agency Health Programme Evaluations in Humanitarian Crises.* WHO/UNHCR. Unpublished.

De Klerk, T and T Nourse (2004) 'Developing Micro-enterprise Refugee Camps: ARC's Experience in West Africa', in *Humanitarian Exchange* No 28 (November).

JEEAR (1996) *The International Response to Conflict and Genocide: Lessons from the Rwanda Experience Copenhagen.* Denmark: Steering Committee of the Joint Evaluation of Emergency Assistance to Rwanda. March.

Levine, S and C Chastre (2004) *Missing the Point: an Analysis of Food Security Interventions in the Great Lakes.* Overseas Development Institute Humanitarian Policy Group Network Paper 47, July. London: ODI.

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

OECD/DAC (1999) *Guidance for Evaluating Humanitarian Assistance in Complex Emergencies.* DAC Working Party on Aid Evaluation.

OECD/DAC (2000) *Effective Practices in Conducting a Multi-Donor Evaluation.* OECD/ DAC Evaluation and Aid Effectiveness Paper 4. OECD/DAC.

OECD/DAC Network on Development Evaluation (2005) *Challenging the Conventional Wisdom – the View from Developing Country Partners*. Workshop Report, Nairobi, 20–24 April 2005.

Poletti, T (2004) *Inter-Agency Health Evaluations in Humanitarian Crises – a Background Issues Paper.* WHO/London School of Hygiene and Tropical Medicine/ UNHCR.

Slim, H and E Eguren (2004) *Humanitarian Protection: A Guidance Booklet, Pilot Version* London: ALNAP/ODI, January

UNHCR and SCF (2002) *Note for Implementing and Operational Partners by UNHCR and Save the Children-UK on Sexual Violence & Exploitation: the Experience of Refugee Children in Guinea, Liberia and Sierra Leone based on Initial Findings and Recommendations from Assessment Mission 22 October–30 November 2001,* February.

WFP (2005) *Summary Report on WFP Follow-up to Evaluation Recommendations.* WFP/ EB.A/2005/7-B. 25 April.

WHO/London School of Hygiene and Tropical Medicine/UNHCR (2004) *Proceedings of the Inter-Agency Health Evaluations in Humanitarian Crises Workshop.* Geneva, 8–9 March 2004.