Section M

Meta-Evaluation

Overview M1

The purpose of a meta-evaluation is to assess the quality of a thematic group of evaluations, and subsequently to suggest improvements. As Lipsey points out (2000:212):

> Meta-analysis and other forms of systematic synthesis of evaluation studies provide the information resources for a continuous improvement program for evaluation practice itself. By examining the patterns and relationships revealed by meta-analysis, an evaluator will better understand what program characteristics, outcome domains, and research methods are most likely to be important for a particular evaluation effort. As new evaluation studies are completed and added to cumulative syntheses, the knowledge resources of the evaluation field will become richer and more differentiated and their potential contribution to practice, in turn, will become more useful.

As in the two previous Annual Reviews, analysis is based on an assessment of the preceeding year's set of evaluations against the ALNAP Quality Proforma (QP), which this year involved 34 English language reports and five French language reports.

Evaluation of Humanitarian Action (EHA) has many strengths. In particular, it allows us to look closely at key management issues, assess performance against some of the DAC criteria (including effectiveness, sustainability/connectedness and relevance/appropriateness), and understand better the contextual background of a particular emergency. This year good practice was found in particular in the reports by WFP and the Disasters Emergency Committee (DEC). While it is difficult to draw definite conclusions there has certainly been an improvement in the focus on longer term results. This is illustrated by a solid minority of reports achieving good rating against the DAC criteria 'impact'. However, this year's analysis also points to five generic weaknesses:

1 failure to use agency policies for evaluation purposes;

2 lack of attention to rights-based approaches (including gender equality) and

protection (evaluators clearly feel more comfortable dealing with management issues or the effectiveness of sectoral interventions than with rights-based issues);

3 questionable credibility of many reports due to inadequate methodology and/or because it is unclear from where conclusions are drawn;

4 failure to consult with primary stakeholders and/or to adequately describe the nature of this consultation;

5 recommendations that are poorly developed and therefore unlikely to be followed.

The last three points may be reasons, among others, why evaluation results are not being picked up on more fully.

This section begins with an introduction to the revised ALNAP QP (reproduced at the end of this chapter) and the assessment process. It then presents findings as organised by the QP headings. There is a general conclusion that, along with Box M1, elucidates a suggested new approach to EHA as recommended in the report on DEC agency interventions after the Gujarat earthquake (DEC, 2001). A year-on-year cumulative comparison for the three years covered by the ALNAP Annual Reviews is included as Box M3.

The ALNAP Quality Proforma (QP) M1.1

The ALNAP QP was developed in 2000 by drawing on what was commonly accepted as good practice in EHA and evaluation in general. It is a 'live' document and continues to evolve as EHA evolves. The QP is not used to rank evaluation reports and no composite rating of individual reports is provided. Rather, the intention is to reach general conclusions on trends as well as strengths and weaknesses in EHA. Assessments using the QP are made entirely on the basis of information contained in an evaluation report; issues related to recommendations and follow-up are not covered unless discussed specifically in the report. As in the previous two years a 'satisfactory' rating is taken as the benchmark for adequate performance (as set out in the Guidance Notes column in the QP).

In the light of previous use in the Annual Review, the QP was revised this year to strengthen analysis. The main revisions were as follows:

- reorganisation for ease of use;
- addition of a section dealing with planning and implementation;
- addition of new Areas of Enquiry relating to: the cost of the evaluation (1.i); evaluator bias (2.v); and protecting confidentiality and promoting respect for stakeholders' dignity and self-worth (2.vi) (the latter areas are included in the American Evaluation Association The Program Evaluation Standards, 1994, and all three areas should be seen as central elements in EHA);
- separate ratings for each of the DAC/OECD criteria to facilitate a disaggregated understanding of evaluation performance against each criterion;
- clarification of some guidance notes, including addition of information in several areas to support determination of what can be considered 'satisfactory'; and removal of the C+ rating ('close to satisfactory') which was considered no longer necessary.

These clarifications and revisions have 'raised the bar'; the requirements to achieve a satisfactory rating have been made more stringent as QP guidance notes have been clarified. Some of the decline in reports' performance this year may be accounted for by these changes. This is noted where relevant in the text.[1]

Assessment Process M1.2

In order to increase rigour and counter the potential for assessor bias and error, the assessments were undertaken by two assessors: the author of this chapter and Peter Wiles. Both were involved in the last two meta-evaluation exercises and the subsequent QP revision. Reports in French were rated by one assessor, Sylvie Robert.[2]

The assessment process for the reports in English was twofold. An initial assessment of the core evaluation reports was undertaken independently by each assessor.

Discussion on issues of interpretation of guidance notes, possible errors and omissions ensued, and was followed by a final independent review by each assessor. The resulting 90 per cent consistency rate was deemed an acceptable margin for the purposes of this meta-evaluation. Where there was inconsistent rating, results are not included in the analysis.[3]

Quality Proforma Follow-up M1.3

The quality assessments completed for Annual Review 2002 were returned to relevant agencies, but elicited almost no response. Thus a more proactive approach has been adopted this year whereby one of the assessors will be available to discuss individual assessments with commissioning agencies.

The Sample M1.4

Thirty-seven reports cover 24 countries or regions. From Africa: Angola (2 reports); Burundi (4 reports); DR Congo (4 reports); Ethiopia (2 reports); Great Lakes; Kenya; Liberia; Mozambique; Sierra Leone (4 reports); Somalia; Sudan; and Uganda. From Central Europe/Asia: Azerbaijan; Afghanistan (3 reports); Bosnia; northern Caucasus; Croatia; Kosovo; and Iran. From Asia: Bangladesh; India and DPR Korea. From Latin America: Brazil and Colombia. The remaining two reports had a multi-country focus (see Figure M1).

Figure M1 Regional Coverage of Core Sample

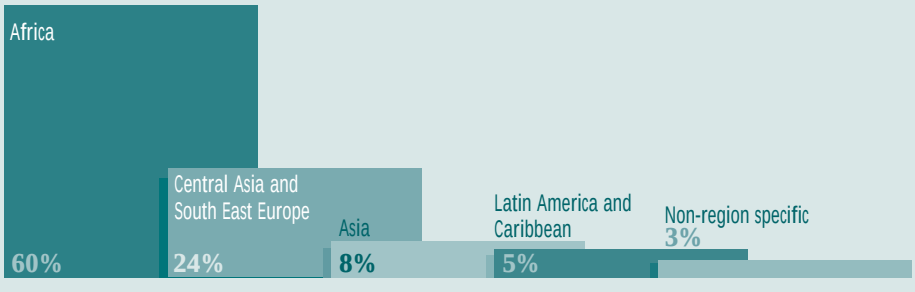
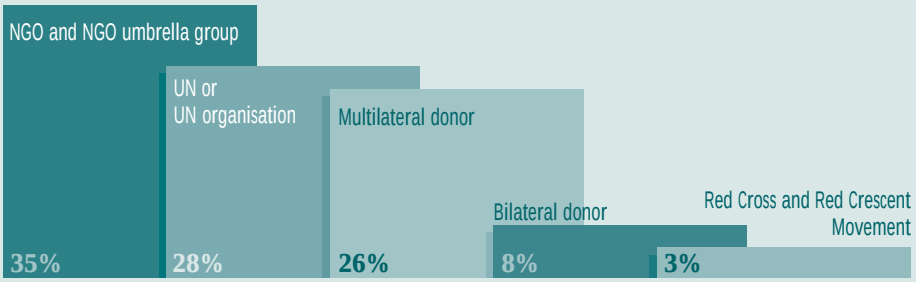


Figure M2 Breakdown of Sample by Type of Commissioning Organisation



Fourteen of the reports were commissioned by NGOs (of which five were reports on Oxfam); 11 by the UN system (of which eight were on WFP); 10 by ECHO; three by bilaterals; and one by the IFRC (see Figure M2).

Timeliness M1.5

In order to facilitate the use of results and recommendations, an analysis was carried out of whether evaluations were completed and reports published in a timely fashion. Thirty-six reports provided data sufficient to facilitate this analysis, although this data was in many cases incomplete or ambiguous. For example, many reports did not specify when the intervention began and ended or the phase of the intervention being evaluated.

The following illustrates that EHA is being carried out in a timely manner: 16 evaluations were conducted on ongoing interventions; 12 evaluations were conducted within one month of completion of operations; four evaluations were conducted within two to three months after completion of operations; and two evaluations were conducted more than three months after completion of operations. The large number of evaluations carried out during ongoing operations is probably due to the long-term nature of many of the interventions included in the 2002 set.

For the most part reports were produced in a timely fashion, with 13 reports finalised within one month of the evaluation, 11 within two to three months, and eight within four to seven months. While we do not have comparative figures for

other sectors, this performance seems at least satisfactory as far as promoting the use of reports is concerned.

Three-year Comparative Analysis M1.6

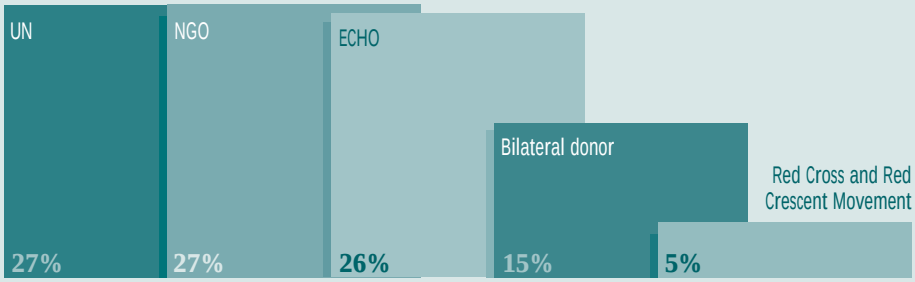
A new feature of this year's meta-evaluation is a year-on-year comparative analysis based on comparable Areas of Enquiry from the QP used in 2001, 2002 and 2003. This comparative analysis involves a total of 127 reports.

As a slightly different rating system has been used over the three years, analysis has been carried out using 'satisfactory' and 'unsatisfactory' ratings only – as defined in the Guidance Notes of the QP.

The breakdown of the sample by agency over three years is given in Figure M3. Between them the UN, NGOs and ECHO constitute 77 per cent of the evaluations included in the meta-analysis in the Annual Reviews over this and the last two years.

The limited number of evaluations commissioned by bilaterals provided to ALNAP each year suggests that Collinson & Buchanan-Smith's (2002) analysis concerning lack of accountability of donors vis-a-vis their willingness to commission independent evaluations holds true. ICRC, with two evaluations provided over three years, is also poorly represented.[4]

Figure M3 Breakdown of Agencies Included in QP Assessment for 2001, 2002 and 2003



While the analysis of EHA shows some strengths, for example, in application of some of the DAC criteria, attention to rights-based issues and consultation with and participation of primary stakeholders stand out as particular weaknesses.

Assessment Against the ALNAP Quality Proforma: 2002 Reports M2

Information included in the Area of Enquiry column in the tables below provides the outline of the area being considered. Further details as to how the rating was determined can be found in the Guidance Notes column in the QP.

Proforma Section 1: Evaluation Terms of Reference (TOR) M2.1

Note: Of the 39 reports assessed, 11 included no TOR in the version received. The analysis below and figures in the tables are therefore based on the 28 reports that included TOR, except for QP area 1.2i.

TOR Focus and Use (1.1ii; 1.1iii; 1.1iv; & 1.1vi)

In terms of the statement on the intervention to be evaluated, reports were required to include adequate details on the emergency context, intervention objectives and key stakeholders involved. Only nine reports that included a TOR managed to provide adequate information in all three areas. Of the remaining 19, details concerning key stakeholders were least often provided (in only four of the 19 reports); better information was included on context and objectives (provided in nine of the 19). Failure to set out clearly in the TOR how primary stakeholders should be consulted may be one of the reasons for poor consultation with and participation of primary stakeholders in the evaluation process (see M2.6 later).

Commissioning agencies should, as a matter of course, include in the TOR a requirement that evaluators consult adequately with primary stakeholders. What can beconsidered 'adequate' is:

Table M1 TOR Focus and Use

Area of Enquiry	Rating	% of Reports Attaining Rating (rounded)
1.1ii		
Quality of TOR statement on the intervention to be evaluated.	Good	4
	Satisfactory	32
	Unsatisfactory	56
	Poor	8
1.1iii		
Quality of TOR statement on the purpose of the evaluation.	Good	26
	Satisfactory	52
	Unsatisfactory	8
	Poor	14
1.1iv		
Quality of TOR statement on the primary focus of the evaluation.	Good	26
	Satisfactory	70
	Unsatisfactory	4
	Poor	0
1.1vi		
Quality of TOR statement on intended use and user(s) of the evaluation output(s).	Good	4
	Satisfactory	11
	Unsatisfactory	63
	Poor	22

- that sufficient information is gained from primary stakeholders, including from both sexes and different ethnic groups, to allow conclusions to be formulated about the intervention;
- that primary stakeholders be given an opportunity to be active participants in the evaluation process, even if only through focus groups or PRA exercises;
- that primary stakeholders perspectives' can be triangulated with those of other key stakeholders.

Reports were strong on providing background information on the purpose and primary focus of the evaluation. Of the 26 reports that covered this area, 17 noted a joint lesson learning and accountability focus, six a lesson learning and three an accountability focus – though for the most part reports did not specify the relative emphasis placed on each. The constraints to achieving both lesson learning and accountability functions from the same evaluation are discussed in Annual Report 2002. It is therefore possible to conclude that commissioning agencies appear to be including both purposes in a rote manner without considering the consequences of this for evaluation process and use. Of the 23 reports noting primary focus, nine had a programme focus (mainly ECHO reports), five had a project focus[5], and the remainder had either a policy, institutional or joint focus.

One of the weaker areas of the reports (1.1vi) is the extent to which they outline the intended uses of the evaluation findings. Sixty-three per cent of reports rated as unsatisfactory and 22 per cent as poor in this area, with no mention of this topic. Furthermore, commissioning agencies are not following up on their responsibility to ensure that evaluation results are used, despite widespread acknowledgement in many cases that reports do not receive sufficient attention. The one good practice example this year is the WFP evaluation of its intervention in the Great Lakes region (September 2002, Annex 1) which notes that: '[T]he report will be presented to WFP's Executive Board; key recommendations arising from the evaluation will be used in the preparation of a Management Response Matrix which will outline how the WFP Regional Bureau in Kampala intends to follow up on the evaluation's key findings and recommendations; and [there will be] dissemination through WFP's website and a publicly available summary.'

TOR process and team make-up: evaluation cost (1.1i)

This is an area where there has been surprisingly little analysis in EHA, and where no interagency standards exist. There does not even appear to be a requirement for commissioning agencies to report on likely costs of evaluations in the TOR. Thus a new Area of Enquiry was added this year, and it was found that only two reports included the cost of the evaluation (Norway Ministry of Foreign Affairs, November 2001; WHO, December 2002). In the former case the cost was some US$175,000; in the latter it was US$20–30,000 out of a total expenditure of US$1.7m in 2002 – or between some one-and-a-half and 2 per cent of total WHO expenditure. This area should be included in TOR for transparency and cost-effectiveness reasons, and to allow an assessment of whether agencies are

Table M2 TOR Process and Team Make-Up

Area of Enquiry	Rating	% of Reports Attaining Rating (rounded)
1.1i		
Cost of the evaluation.	Good	4
	Satisfactory	4
	Unsatisfactory	0
	Poor	92
1.1v		
Quality of TOR statement on the expectation of good practice in approach and method.	Good	4
	Satisfactory	0
	Unsatisfactory	86
	Poor	11
1.1vii		
Quality of TOR guidance on the evaluation report format.	Good	29
	Satisfactory	11
	Unsatisfactory	7
	Poor	54
1.1viii		
Evaluation Timeframe *a. Timeliness* The TOR should outline the rationale for the timing of the evaluation.	Good	15
	Satisfactory	25
	Unsatisfactory	0
	Poor	60
1.1viii		
Evaluation Timeframe *b. Sufficiency* Sufficient time should be allowed to develop methods; review background / contextual information; carry out fieldwork; undertake analysis at all stages of the evaluation; and finalise the report.	Good	0
	Satisfactory	50
	Unsatisfactory	21
	Poor	29
1.1ix		
Quality of TOR clarification process.	Good	4
	Satisfactory	11
	Unsatisfactory	4
	Poor	81
1.2i		
Nature, make-up and appropriateness of the evaluation team.	Good	8
	Satisfactory	8
	Unsatisfactory	38
	Poor	46

spending an adequate percentage of an intervention budget on evaluation where agency standards do exist.[6]

Expectation of good practice in approach and method (1.1v)

In this area, TOR are expected to outline application of DAC criteria; reference international standards, including international law; talk about the importance of a multi-method approach; explain the consultation process with key stakeholders, including primary stakeholders; and bring in the key issue of gender analysis. Only one report, the DEC evaluation of interventions after the Gujarat earthquake (DEC, December 2001), managed to cover all these areas adequately.

It is useful to consider the disaggregated breakdown of this section. Almost no TOR required evaluators to either use international standards, such as the Sphere standards, or to examine whether the intervention had used international standards during its implementation. Protection was similarly largely ignored in the TOR as an issue to be evaluated. On the other hand, TOR generally set out a requirement to use the DAC criteria and explained what the criteria meant (in 19 out of 28 reports); a majority of TOR also required attention to gender equality (16 out of 28 reports). The requirement to develop a multi-method approach and consult with key stakeholders was weaker (11 out of 28 reports).

In the light of this disaggregated breakdown, there is a fairly clear correlation between requirements in the TOR and what was actually done by evaluators, particularly in terms of the lack of attention to international standards and protection and the relative success in application of the DAC criteria. Commissioning agencies and evaluators may want to mull this over when formulating and finalising their TOR. The area of clarification of TOR between the commissioning agency and evaluation team (1.1ix) is one that we hear almost nothing about in evaluation reports, with 81 per cent of reports with TOR making no mention of this topic despite its importance to the overall direction of the evaluation.

Evaluation timeframe (1.viii.a & b)

The evaluation timeframe is also inadequately reported in terms of the reason why the evaluation is being carried out at a particular time (60 per cent of evaluations rated as poor); however, reporting on whether sufficient time had been allowed for the evaluation was better (50 per cent rated as satisfactory).

The nature, make-up and appropriateness of the evaluation team (1.2i)

This area is important for establishing the credibility of the evaluation process, yet only four reports managed to do this (DEC, December 2001; Norway Ministry of Foreign Affairs, November 2001; Oxfam, September 2001; and WHO, December 2002). Forty-six per cent of the 28 reports with TOR provided no information on this area at all.

Of particular interest here is the make-up of evaluation teams which is included in, or could be extrapolated from, reports. Of the 61 evaluators employed, 46 were expatriates, seven were locally based consultants (usually citizens of the countries where the evaluation took place), and eight were agency staff members (from two Oxfam evaluations). While this is a better balance than that noted in Annual Review 2001 for the Kosovo evaluations, where 52 of 55 consultants were expatriates, this still represents a serious imbalance and under-employment of locally based professionals. Only the DEC has evidenced consistent good practice in this area, this year hiring a team with a complementary mix of international and locally based consultants (DEC, December 2001). In addition, of the 61 evaluators, 10 were internal agency staff. However, the implications of this for evaluation practice are not discussed.

One of the key cross-cutting themes in Chapter 3 this year was the general lack of capacity building in humanitarian action. The same could be said for EHA; agencies could usefully maintain and use rosters of locally based evaluators and attempt to ensure a better mix of international and locally based evaluators. ALNAP could also play a central role here, both in terms of developing national evaluation capacity through training and maintaining a roster of consultants.[7]

Proforma Section 2: Evaluation Approach and Methods M2.2

Overview

Description of the evaluation approach did not display any systematic good practice, and report methodology sections in general did not provide a basis from which it was possible to assess the likely accuracy of findings. This may be because evaluators do not feel it is necessary to elaborate on the methodology used. However, this lack is specific to EHA; in both mainstream evaluation practice and the evaluation of development interventions, greater attention is given to establishing the credibility

Table M3 Appropriateness of Evaluation Approach

Area of Enquiry	Rating	% of Reports Attaining Rating (rounded)
2.i Appropriateness of the overall evaluation approach.	Good	0
	Satisfactory	5
	Unsatisfactory	8
	Poor	87
2.ii Appropriateness of the evaluation methods selected.	Good	3
	Satisfactory	0
	Unsatisfactory	69
	Poor	28
2.iii Appropriateness of the planned application of the DAC criteria and rationale.	Good	5
	Satisfactory	8
	Unsatisfactory	13
	Poor	74
2.iv Consideration given to constraints.	Good	0
	Satisfactory	19
	Unsatisfactory	42
	Poor	39
2.v Consideration given to evaluator bias.	Good	0
	Satisfactory	3
	Unsatisfactory	5
	Poor	92
2.vi Consideration given to confidentiality and dignity.	Good	3
	Satisfactory	3
	Unsatisfactory	7
	Poor	87
5.3i Quality of application of the selected evaluation methods.	Good	3
	Satisfactory	5
	Unsatisfactory	76
	Poor	16

of evaluation methodology. Experienced evaluators know that when evaluations bring unwelcome findings and recommendations the first thing that may be questioned is the evaluation methodology.

Producing a credible description of methodology would, in the majority of cases, require only a little more effort – for example, noting the numbers of primary stakeholders consulted, broken down by sex and other salient social characteristics; where and when they were consulted; and the methods used for consultation (survey questionnaire, focus group, etc). That the majority of reports do not report even such basic information suggests sloppy practice on the part of evaluators and commissioning agencies, and undermines their credibility.

Appropriateness of the overall evaluation approach (2.i)

Reports were assessed concerning the extent to which the overall evaluation approach was clearly outlined and the appropriateness of choice established relative to the evaluation's primary purpose, focus and end-users. 'Approach' here means the wider conceptual framework used and the evaluation tradition being drawn upon, such as accountability oriented, utilisation-focused, or empowerment evaluation approaches.

Only two reports were assessed as satisfactory in this area: the evaluation of the DEC's intervention after the Gujarat earthquake (DEC, December 2001) and the evaluation of the Norwegian Red Cross (Norwegian Ministry of Foreign Affairs, November 2001). Eighty-seven per cent of the set rated as poor. In the DEC case the rationale for primary stakeholder consultation is clearly set out and the community survey technique used is close to the empowerment evaluation approach. The Norwegian Red Cross evaluation analyses the way in which definition of the term 'humanitarian' affected the evaluation methodology. In both cases there was an explicit discussion of why a particular evaluation approach was taken.

The implications of the lack of attention to wider evaluation discourse were discussed in detail in Annual Review 2002, and include: a lack of conceptual direction for the evaluation; an inability to rationalise why a particular evaluation methodology has been selected; a fall-back on 'standard' evaluation techniques with little experimentation; and lack of attention to causality.

Appropriateness of the evaluation methods selected and appropriateness of planned application of the DAC criteria (2.ii & iii)

A large majority of reports were rated as unsatisfactory or poor in both these areas. In terms of rating of appropriateness of selected evaluation methods, reports most often missed reference to international standards, including international law, and gender analysis. Among those reports rated satisfactory or better was the evaluation of WFP's intervention in Iran supporting Iraqi and Afghan refugees (WFP, September 2002b). This report discusses in some detail the reason for selection of particular camps to visit and the random selection of refugees for interview purposes. It provides a checklist for refugee camp visits and notes the methods used – in particular, focus groups. It is also gender sensitive and notes the use of control groups.

Use of control groups is a rare phenomenon in EHA, and it is common for reports to argue that they cannot conclude whether results were caused by the intervention because of difficulties of attribution. For example, World Vision (June 2001:5) notes: '[W]hile the emergency interventions undoubtedly played an important role in reducing the levels of malnutrition, a number of other factors may have contributed to this reduction.' Use of a rudimentary control group approach, for example, interviews with a small sample of the affected population who have not been included in the intervention target group, can help overcome this problem. Not surprisingly, because a majority of reports were assessed as unsatisfactory in selection and detailing of evaluation methodology a similar number were assessed as unsatisfactory in the use of evaluation methods (5.3i).

The UNHCR (May 2002) evaluation of the protection of children, not included in this year's meta-analysis because of its thematic focus, is in many respects an example of good practice in the application of methodology. In particular, the report has included an Annex on lessons learned in the evaluation process and methodology. These lessons include: the need for triangulation between different sources of data; the need for the UNHCR evaluation office to communicate the important balance between 'independence' and 'internal purview' when announcing the evaluation; the importance of having a representative and active steering committee to guide the evaluation; the importance of having briefing and debriefing sessions with the country office; and the key role that focus groups can play in providing qualitative information.

In terms of the application of DAC criteria, reports were required to address each of seven criteria to achieve a satisfactory rating. Seventy-four per cent of reports did not discuss the criteria at all in the methods section, except in a very general fashion and even where attention to the criteria was included as a requirement in the TOR. Of those that did include a discussion, 'coherence' was the criterion most often missed.[8]

Consideration given to evaluator bias; and stakeholder confidentiality and dignity (2.v & 2.vi)

Bias is a well-known feature of evaluations, both conceptual bias that inevitably comes with any individual's perspective and/or bias that is introduced as a result of an evaluator having been associated with an intervention. One of the key points of a well-designed methodology is to guard against bias. Evaluations did not see fit to address this topic. For example, seven evaluations involved staff from the agency being evaluated as part of the evaluation team. However, there is no discussion in the reports as to why a staff member was included and, for example, the potential bias or benefits this brings.

Ensuring confidentiality and the dignity of key stakeholders should be central to any evaluation, but is particularly important in EHA where primary stakeholders may be at risk and where they have often been subject to trauma. Doubtless evaluators treat primary stakeholders with respect and would never dream of putting them at risk intentionally; it is important to acknowledge this in their reports. In a similar fashion, evaluation reports should point out how the views of other key stakeholders are kept confidential and how the evaluation method encouraged key stakeholders to express their independent opinions. Ethical research standards for interviews do exist, for example, providing interviewees with a form signed by the evaluators noting that their views will be kept confidential. Such a system could be usefully adapted for EHA.

An evaluation may be the first time primary stakeholders are listened to seriously (DEC, December 2001, Vol 3, Methodology: pt 6): 'Some members of the community stated that no one else had asked what they wanted or needed, or how they felt about the response.' This raises the question as to whether part of the purpose of an evaluation should be to give voice to the usually voiceless – as professed in empowerment evaluation and as the DEC report argues. Not all evaluators would agree with this perspective but rather consider that evaluation should be an

'objective' exercise. This is why providing a rationale for the evaluation approach and details on potential evaluator bias is crucial.

Assessing Contextual Analysis M2.3

Overview

The context section of an evaluation report should constitute a concise and relevant background that allows the reader to understand the situation in which the intervention took place, and how context is relevant to, and has been taken into consideration in, the evaluation. Understanding of context is important for evaluation purposes primarily because it supports attribution of results. This year's reports rated fairly well in attention to context. To receive a satisfactory rating reports were required to include relevant details on historical, social (including gender analysis), economic, political, and cultural features. Reports tended to focus on providing background to the sector being assessed, and most often left out historical and cultural features.

Analysis of context (3.i)

In terms of analysis of the affected area and population there was significant good practice, for example USAID (April 2001), WFP (January 2002; April 2002) and ECHO (October 2001b). The WFP Angola evaluation (April 2002) section on context includes details on the war and agriculture and the impact of this on food aid and the food economy; coping strategies of primary stakeholders (including trade in semi-urban settlements); the economy; geographical location of insecurity (including a map); gender and poverty; land tenure; and the lack of government policy on LRRD. All of these areas are of relevance to the intervention and are drawn upon in the analysis of results.

Quality of use of context information (5.1i)

Fewer reports integrated the discussion of context into the analysis, and 63 per cent of reports were rated as unsatisfactory or worse in this area - although the tendency to provide stand-alone contextual sections with little reference to the rest of the report had diminished. On the other hand about 80 per cent of those reports that did include a satisfactory or better context section also managed to integrate this with discussion of results – i.e., to demonstrate how context affected achievement.

Table M4 Quality of Contextual Analysis

Area of Enquiry	Rating	% of Reports Attaining Rating (rounded)
3.i		
Quality of the evaluation's analysis of context.	Good	16
	Satisfactory	32
	Unsatisfactory	39
	Poor	13
3.ii		
Quality of the evaluation's analysis of past involvement of the agency and its local partners.	Good	7
	Satisfactory	29
	Unsatisfactory	32
	Poor	32
3.iii		
Quality of the evaluation's analysis of the crisis to which the intervention is responding.	Good	9
	Satisfactory	15
	Unsatisfactory	43
	Poor	33
5.1i		
Quality of the use made by the evaluation of contextual information.	Good	14
	Satisfactory	23
	Unsatisfactory	49
	Poor	14

Agency involvement (3.ii)

Details of the past involvement of the agency and its partners in the geographical area of the intervention is less well covered. Indeed it might appear that the intervention occurred in a historical vacuum in many of the reports, whereas the prior involvement of an agency in an area has been identified as a key factor in success (see Annual Review 2002). That one-quarter of evaluation reports did not see fit to cover this area, and a further 43 per cent were rated as unsatisfactory, suggests that the importance of prior involvement and building partnerships which can be drawn on in an emergency is not sufficiently recognised in EHA.

Assessing the Analysis of the Intervention M2.4

Evaluation of policies and principles (4.1i)

Reports were generally weak in terms of evaluating adherence to policies. Reports commissioned by ECHO and NGOs tended to be weaker in this area, with only one report for each type of agency rating as satisfactory (ECHO, December 2001f; DEC, December 2001). Reports commissioned by UN agencies were stronger, with four (three WFP and one UNMAS, February 2002) reports rated as good, and two (both WFP) as satisfactory. Each of these WFP reports tells the story of how the introduction of the WFP policy From Crisis to Recovery influenced the planning and implementation of the respective country programmes; the Great Lakes report (September 2002) goes further and analyses the relation between government policy and WFP's intervention.

It is surprising that more evaluations do not use agency policy as a standard against which results can be measured given the key role that such policy should play in guiding agency action. This is partly explained by TOR not including this as a requirement.

Table M5 Instiutional Considerations

Area of Enquiry	Rating	% of Reports Attaining Rating (rounded)
4.1i		
Quality of the evaluation of agency guiding policies and principles.	Good	12
	Satisfactory	17
	Unsatisfactory	21
	Poor	50
4.1ii		
Quality of the evaluation of an agency's management and human resource practices.	Good	18
	Satisfactory	37
	Unsatisfactory	30
	Poor	15

Management and human resource practices (4.1ii)

As in the two previous Annual Reviews, this area has proven to be a strength with 55 per cent of reports this year rating as satisfactory or good. Reports contained most information about staff turnover and field/HQ relations; less information was provided on briefing and debriefing procedures and training.

There were a number of examples of good practice where reports made the connection between management practices and intervention results. Among these was the assessment of CARE's programme in Afghanistan (CARE, September 2002) which includes extensive analysis of field/HQ communication; the implications of a lack of training on intervention impact; lack of learning from earlier emergencies (e.g., around the importance of income generation schemes in the recovery phase); and how security issues affected programming. Another good practice case was the evaluation of the IFRC intervention after the Goma volcanic eruption (IFRC, September 2002) This provided a good assessment of intra-agency communication/coordination, including dispatch of emergency teams; intra-Federation communication; level of preparedness of Federation staff; the level of experience of emergency teams; and training.

Overview

This section of the QP was one of those areas revised this year to reflect more accurately the various stages of the project cycle, from planning to monitoring. Overall, evaluation of the project cycle process was a relatively strong area of assessment with reports rated as satisfactory or better in about 50 per cent in three areas: evaluation of implementation, monitoring, and expenditure. Reports tended to be consistent in their coverage of project planning and implementation – that is, they were either rated as satisfactory or better, or unsatisfactory or worse, in all areas of enquiry. WFP reports were particularly strong in this area.

Evaluation of needs and livelihoods assessment (4.2i)

Forty per cent of reports were attuned to this issue and included an analysis of both whether the intervention had carried out an adequate needs assessment, and the importance of this. Several reports (e.g., World Vision, June 2001; DEC, December 2001; WFP, September 2002b) go beyond a critique of the lack of an adequate needs assessment and include an analysis of the ways in which livelihood strategies could have been more adequately covered in the intervention planning process.

Table M6 Needs Assessment, Objectives, Planning and Implementation

Area of Enquiry	Rating	% of Reports Attaining Rating (rounded)
4.2i		
Quality of evaluation of the needs and livelihoods assessments(s) that informed the intervention.	Good	8
	Satisfactory	32
	Unsatisfactory	30
	Poor	30
4.2ii		
Quality of evaluation of the intervention objective(s).	Good	22
	Satisfactory	19
	Unsatisfactory	47
	Poor	12
4.2iii		
Quality of evaluation of the intervention planning processes (including design).	Good	11
	Satisfactory	22
	Unsatisfactory	62
	Poor	5
4.2iv		
Quality of evaluation of the intervention implementation process.	Good	10
	Satisfactory	39
	Unsatisfactory	51
	Poor	0
4.2va		
Quality of evaluation of monitoring and/or real-time evaluation mechanisms. • Analysis of the intervention's monitoring and/or real-time evaluation mechanisms and the effect on intervention results.	Good	19
	Satisfactory	44
	Unsatisfactory	25
	Poor	12
4.2vb		
Quality of evaluation of monitoring and/or real-time evaluation mechanisms. • Assessment of the indicators used. Where the intervention activities span relief, rehabilitation and/or development, indicators should be evaluated relative to each type of activity.	Good	0
	Satisfactory	41
	Unsatisfactory	24
	Poor	35
4.2vi		
Quality of evaluation of the intervention expenditure.	Good	18
	Satisfactory	32
	Unsatisfactory	32
	Poor	18

For those reports rated as unsatisfactory the majority either only mentioned the livelihoods assessment in passing or included no details on primary stakeholder consultation and participation. Almost one-third of reports made no mention of this topic.

Evaluation of intervention planning processes (4.2iii)

This has proven a difficult area for EHA to come to grips with, with some 67 per cent of evaluations (25 reports) rated as unsatisfactory or poor. Of these 25 reports, 19 did not include adequate details on primary stakeholder consultation and participation in the planning processes; as was seen in Chapter 3, facilitating consultation and participation in planning is one of the most difficult areas of humanitarian action. By failing to pay systematic attention to this issue evaluators are compounding the problem.

Evaluation of the intervention implementation processes (4.2iv)

Evaluation of implementation processes was satisfactory in about half of all reports. However, lack of attention to primary stakeholder participation and consultation was a significant problem. UN agencies rated higher than NGOs, suggesting that the perceived greater capacity of NGOs to foster consultation and participation of primary stakeholders has not translated into capacity to facilitate evaluation of this aspect of humanitarian action.

Evaluation of monitoring and/or RTE mechanisms, and indicators (4.2va & b)

Sixty-three per cent of reports in the former area were assessed as satisfactory or better. Of the six reports rated as good in relation to evaluation of monitoring, four were commissioned by WFP (April 2002, September 2002, September 2002b, December 2001), one by the Norwegian Ministry of Foreign Affairs (November 2001), and one by ECHO (December 2001c). The significant number of reports that did not cover monitoring adequately is surprising given that this is usually an area close to evaluators' hearts. That four reports barely touched on this issue suggests a major oversight by commissioning offices in their vetting of reports. In terms of attention to indicators, 59 per cent of reports did not cover this area even though the development of indicators is essential to results-based planning and should be a central feature of evaluators' approaches.

Quality of evaluation of the intervention expenditure (4.2vi)

Given the arcane budget codes of some agencies it is often difficult to assess whether funds are used for relief or rehabilitation – phases of a response which in any case

often merge into each other. As noted in the previous two Annual Reviews, agencies need to do a better job of delineating the different stages of their interventions, including financial allocations, so that relief and rehabilitation can be evaluated against appropriate indicators.

That 50 per cent of reports were assessed as good or satisfactory in this area would appear to be an improvement on the past two years, although because of differences in phrasing in the Proforma it is not possible to make direct comparison. Of the six examples of good practice, four were evaluations commissioned by WFP and two by NGOs. It is not surprising to find a sound level of detail in the WFP reports as part of their mandate was to investigate the LRRD continuum. In total, about half of the ECHO and NGO reports performed credibly.

Consideration Given to Cross-cutting Themes M2.5

Evaluation of the intervention's adherence to international standards (4.3i)

As with last year's assessment, evaluation of how far the intervention adhered to international standards was again inadequate, with 86 per cent of reports rated as unsatisfactory or poor (the same as in Annual Review 2002). International standards such as the Sphere standards are either not being used by agencies or, less likely, this issue is not being picked up by evaluators. A parallel finding is that evaluators are not themselves using international standards as a means of evaluating interventions (5.3iii) - one of the reasons for this being that most EHA is organised around the OECD/DAC criteria to the seeming exclusion of other international standards. This in turn may be because most commissioning agencies do not require that these standards be used (see the analysis of Area of Enquiry 1.v on good practice in evaluation method, in Section M2.1 above). There is one innovative good practice example in this area, showcased in Box M1, which is the DEC's evaluation of NGO interventions after the Gujarat earthquake.

Consideration given to coordination activities (4.3ii)

Results were satisfactory or better in 52 per cent of cases – not as positive as for 2002 when the figure was 67 per cent. Five reports were rated as good this year, including the evaluation of WFP interventions in Azerbaijan (April 2002a). This covers WFP's coordination mechanisms and relations with the national government, other UN agencies, and implementing partners. The majority of reports that include

Table M7 Evaluation of Cross-cutting Themes

Area of Enquiry	Rating	% of Reports Attaining Rating (rounded)
4.3i		
Quality of evaluation of the intervention's adherence to international standards.	Good	3
	Satisfactory	11
	Unsatisfactory	35
	Poor	51
4.3ii		
Quality of evaluation of the consideration given to coordination activities.	Good	16
	Satisfactory	36
	Unsatisfactory	29
	Poor	19
4.3iii		
Quality of evaluation of the consideration given to protection.	Good	4
	Satisfactory	4
	Unsatisfactory	24
	Poor	68
4.3iv		
Quality of evaluation of the consideration given to gender equality.	Good	14
	Satisfactory	5
	Unsatisfactory	38
	Poor	43
4.3v		
Quality of evaluation of the consideration given to vulnerable / marginalised groups.	Good	12
	Satisfactory	24
	Unsatisfactory	38
	Poor	26
5.3iii		
Reference made to international standards.	Good	3
	Satisfactory	5
	Unsatisfactory	34
	Poor	58

Box M1 Good Practice in the Use of International Standards? Using the Red Cross / Red Crescent Code of Conduct and Sphere Standards

The DEC evaluation of eight British NGO interventions after the January 2001 Gujarat earthquake included innovative use of the Red Cross/Red Crescent Code of Conduct as a measure of performance (December 2001:6):

> We use the Red Cross Code as the basis from which to explore values because it is the most widely accepted set of humanitarian values and all DEC members must sign up to it … The Code was evolved in the West and has not been negotiated with local NGOs or the people in need. In the decade since the Code was devised little has been done to promote it and too often it is just a 'badge' acquired easily by declaration … but it is in the public domain, and anyone donating to the DEC or receiving its aid could reasonably expect agencies to follow it.

The intervention is then evaluated against the 10 sections of the Code, which are rated on points out of 10, with a cumulative rating given. The evaluation also includes an assessment against the Sphere standards in respect to training and DEC members' awareness of the standards, and the water, sanitation and shelter standards.

Without doubt the Code of Conduct is one standard against which humanitarian action should be assessed. However, the DEC report does not make the case as to why it should be the main standard and therefore why it should replace those evaluation mechanisms that are widely understood and in use, such as the OECD/DAC criteria. Evaluators have come to understand over the last decade that 'paradigm wars' as to the most effective evaluation approach are often not useful, and that the most effective evaluation approaches are those that use complementary methods. In any case there is considerable overlap between the several sections of the Code of Conduct and the OECD/DAC criteria (e.g., in relation to coherence, coverage, appropriateness/relevance and sustainability/connectedness) and most of the OECD/DAC criteria are covered in the DEC report. Lastly, while assigning scores to each of the sections of the Code could be useful, it would need more detailed discussion of how scores are to be assigned and whether any weighting should be given before this could be carried out comparatively.

Nevertheless the DEC innovation is certainly welcome and stands out in a field where there is very little experimentation.

an analysis of coordination issues tend to focus on sharing of knowledge through meetings; the WFP Azerbaijan report goes beyond this to a more detailed level of analysis that includes joint planning and implementation activities, as well as knowledge sharing.

As was found in 2002, reports tend to pay greater attention to interagency coordination; their relative lack of attention to coordination with government and local authorities may be part of the overall poor focus on capacity development in humanitarian action, noted in Chapter 3.

Evaluation of consideration given to protection (4.3iii)

Protection continues to be the cross-cutting area least well covered by EHA. Sixty-eight per cent of reports made no mention of protection, a similar finding to 2002, and 92 per cent were rated as unsatisfactory or poor, a worse performance than 2002 when the equivalent was 79 per cent. This despite a considerably greater number of reports on complex emergencies this year.

One of the reasons protection is so poorly covered is that evaluators may see protection as the exclusive mandate of the ICRC and UNHCR. Also, most evaluations have a fairly narrow sectoral focus and do not tend to look far beyond the 'technical' specifics of the intervention – such as the kinds of food provided and to whom, or how many pumps were sunk and whether they are still functioning. In terms of coverage, most NGO and ECHO reports did not cover this area. Conversely, it is interesting to note that of the WFP reports, and although WFP does not in general advocate a rights-based approach, the evaluation of the WFP Angola intervention (April 2002) integrates a detailed discussion of food-related protection issues, including the need for WFP to develop its programme to maximise protection; analysis of whether to refuse to distribute food in cases of forced displacement; and the potential for the provision of food aid decreasing the security of primary stakeholders. Individual evaluators attuned to protection questions can make recommendations on this issue even if this is not required by the TOR. But once again whether they do or not comes back to the question of how far the evaluator should advocate on controversial issues.

The other report included in the meta-evaluation that covered protection thoroughly is that of the Norwegian Ministry of Foreign Affairs (November 2001) – as might be expected in a report on a national Red Cross organisation. The UNHCR (May 2002) evaluation of its work on the protection of children, not

covered in the meta-evaluation because of its thematic focus, nevertheless provides an example of good practice and could be used by other agencies as an example as to what can be achieved in the evaluation of protection. This report clearly analyses the ways in which agency policies and principles on protection were applied. The methodology for this evaluation includes a strong focus on primary stakeholder consultation, and the evaluation notes that triangulation has been carried out – although the process by which this took place is not evident. While overall the evaluation is very rigorous, at times its conceptual discussion could have been complemented by more fieldbased observations; and because of limited focus on impact it is sometimes difficult to determine why conclusions were drawn, for example, concerning linking social and legal protection foci.

Evaluation of consideration given to gender equality (4.3iv)

Attention to gender equality was rated as less than satisfactory or poor in 81 per cent of cases, and in close to half the reports gender is not even mentioned.[9] Both the EU and NGOs performed badly. In contrast, six of the eight WFP reports were rated as good in this area and one as satisfactory; a considerable achievement given overall agency performance. In addition, WFP published a separate thematic report on its Commitments to Women. This impressive attention to gender equality has already been highlighted in Chapter 3 with reference to the results of WFP's interventions.

In most of the WFP reports attention to gender equality is mainstreamed throughout each report as well as being included in a separate section, the latter often being quite substantial. Also of note is an Annex in the reports which contains a checklist on 'Meeting the WFP Commitments to Women and Mainstreaming a Gender Perspective'. Each of the five commitments and their components are rated on a scale from very high to very low, and the reports include detailed narrative observations to complement the rating. The quality of reporting also suggests that WFP has made a commitment to hiring evaluators who have relevant skills in assessing gender equality. Overall, this is probably the most sustained attention to the evaluation of gender equality in EHA to date.

Consideration given to vulnerable/marginalised groups (4.3v)

In the QP definition, vulnerable and marginalised groups include the elderly, disabled, children, and people with HIV/AIDS. As many agencies have policies that

focus their programmes on the most vulnerable, one would expect substantial attention to these in the evaluation reports.

Despite this the picture is mixed, with only 36 per cent of reports rated as satisfactory or better. A problem identified in Annual Review 2002 – the failure of evaluators to disaggregate primary stakeholders – was also found this year, though to a lesser extent. There may be a hangover here from hiring evaluators who have technical expertise (e.g., water or health specialists) and not complementing this with social science expertise.

Of note is the attention given to HIV/AIDS in the ECHO reports on Burundi (ECHO, December 2001e, f, g). In general those reports that pay adequate attention to gender equality also tend to evidence a satisfactory level of attention to the vulnerable and marginalised, although only two reports were rated as 'good' in both these areas: the WFP reports on Iran and the Great Lakes (September 2002b, 2002). This suggests that a consistent attention to basic areas of social differentiation is hard for evaluators to achieve, and is an area where capacity development and training is needed.

Assessment of Evaluation Practice M2.6

Consultation with and participation by primary stakeholders (5.2i)

In order to achieve a satisfactory rating for this area, reports were required to provide adequate detail on the nature (e.g., focus groups) and scope (e.g., numbers by sex of those consulted) of consultation and participation. The failure to do this, noted in the two previous Annual Reviews, worsened this year. Only four evaluations were considered to have undertaken adequate consultation and describe in sufficient detail the consultation that occurred.

It is clear that many evaluators are talking to primary stakeholders, and some of the reports are peppered with their quotations or comments. Why evaluators do not detail the method behind these interviews adequately may be because:

- they do not see the relevance of including this information in the reports, thinking perhaps that it will lead to information overload;

Table M8 Consultation and Participation During the Evaluation Process

Area of Enquiry	Rating	% of Reports Attaining Rating (rounded)
5.2i		
Quality of consultation with and participation by primary stakeholders (beneficiaries and non-beneficiaries within the affected population) during the evaluation.	Good	5
	Satisfactory	5
	Unsatisfactory	27
	Poor	63
5.2ii		
Quality of consultation with, and participation by, other key stakeholders in the evaluation process.	Good	6
	Satisfactory	8
	Unsatisfactory	66
	Poor	20

- they are not aware of the importance of explicitly comparing the perspectives of different stakeholders to add credibility to the evaluation findings;
- much of the focus of EHA is on intra-institutional matters and field trips to project sites are rushed and given low priority.

However, consistent consultation with primary stakeholders, cross-referencing this with other perspectives and detailing the nature and scope of consultation will go a long way to overcoming one of the principal problems with EHA: its failure in many cases to establish credibility of evaluation methods. It will also help fulfil the participatory mandate of most agencies.

The exceptions this year prove that adequate consultation is possible. Of the four reports that were rated satisfactory or better (DEC, December 2001; WFP, April 2002; Oxfam, March 2002; Netherlands Ministry of Foreign Affairs, January 2002), the outstanding report was the DEC evaluation of British NGOs after the 2001 Gujarat earthquake, highlighted in Box M2. Extended consultation in the evaluation of the WFP intervention in Angola (WFP, April 2002) should also be noted.

Box M2 Good Practice in Consultation with Primary Stakeholders

DEC evaluations have consistently consulted with primary stakeholders (see Annual Review 2002). The evaluation of DEC agencies' performance in their response to the 2001 Gujarat earthquake (December 2001) is an excellent example of the levels of consultation that can be achieved:

- The evaluation notes the importance of attempting to empower communities through evaluation approaches that seek their active participation.
- The evaluation team included an Ahmedabad-based disasters institute, the Disaster Management Institute (DMI). DMI organised and conducted a survey covering 50 villages, and interviews with over 2,300 people. The inclusion of national researchers and consultants is a regular feature of DEC evaluations, unlike most other EHA.
- Interviews and focus groups were carried out using state-of-the-art participatory methodologies, and there was considerable attention paid to the location of consultation exercises in order to encourage the participation of as diverse a cross section of the community as possible.
- Specific attempts were made to include 'missing voices', including low status communities, the poorly educated, widows, women, the disabled and sick, those living on the outskirts of communities and working in nearby towns during the day. Timing and location of exercises and follow-up interviews attempted to include these groups.
- The methodology is detailed extensively.
- Quotes and comments from primary stakeholders are used effectively throughout the report to substantiate key points.

Use of the community survey has some weaknesses – for example, primary stakeholders were asked about the total intervention rather than specifically about the DEC agency intervention and this is not taken into account adequately in the analysis. Furthermore, conclusions in the report are sometimes at odds with the findings of the community survey. But it remains an impressive example of what can be accomplished given local expertise and the belief and willingness of the commissioning agency that such an exercise is worth pursuing.

Consultation with and participation by other key stakeholders (5.2ii)

Although 86 per cent of reports were rated as unsatisfactory or poor in this area, much of this can be attributed to a requirement introduced to this year's QP that requires reports to explain the nature of such consultation (e.g., whether confidentiality was ensured). The rationale for this was that there can be a significant variation in responses dependent on the circumstances of an interview – for example, when a senior officer or other third person is present at the interview. As much of the weight in EHA rests on interviews with agency staff, it was thought important to include an assessment of how far the evaluation team facilitated independent expression of views.

The majority of evaluations did involve significant discussions with key stakeholders. However, they also featured generic problems. These included:

- consultation with only one set of stakeholders, usually agency staff, to the exclusion of national and local governments;
- lack of detail on the nature of the consultation, e.g., where it took place, who was present, or whether a questionnaire was used;
- failure to provide a list of key stakeholders consulted.

As with consultation with primary stakeholders the second and third bullet points could easily be corrected in many cases. Including adequate information on these areas will strengthen the credibility of reports.

Quality of application of standard EHA criteria (5.3ii)

Application of the DAC/OECD criteria is one of the stronger areas of EHA. Reports rated highly in the evaluation of effectiveness, relevance/appropriateness, and sustainability/connectedness, where there is a range of good practice. This suggests that use of these criteria has been mainstreamed into EHA and to a lesser extent into evaluation of coverage. Efficiency, impact and, in particular, coherence, fared less well.

First, the good practice. Sixty-eight per cent of reports were rated as satisfactory or good in their evaluation of relevance/appropriateness. Reports rated as good include DEC (December 2001), WFP (December 2001, January 2002), and CARE

Table M9 Application of Methods, Criteria and Standards

Area of Enquiry 5.3ii Quality of Application of EHA criteria in the assessment of intervention.	Rating	% of Reports Attaining Rating (rounded)
a Efficiency (including cost-effectiveness)	Good	6
	Satisfactory	28
	Unsatisfactory	46
	Poor	20
b Effectiveness (including timeliness)	Good	11
	Satisfactory	63
	Unsatisfactory	23
	Poor	3
c Impact	Good	6
	Satisfactory	29
	Unsatisfactory	41
	Poor	24
d Relevance / appropriateness	Good	17
	Satisfactory	51
	Unsatisfactory	23
	Poor	9
e Sustainability / connectedness	Good	14
	Satisfactory	59
	Unsatisfactory	22
	Poor	5
f Coverage	Good	21
	Satisfactory	35
	Unsatisfactory	26
	Poor	18
g Coherence	Good	0
	Satisfactory	12
	Unsatisfactory	15
	Poor	73

(September 2002). The report on WFP's interventions in Somalia, for example, analysed the overall strategy of delivery of food aid (included in the discussion of food aid in Chapter 3), and integrated considerable detail about the appropriateness of the ration.

Seventy-four per cent of reports were rated as satisfactory or better in their assessment of effectiveness and, in particular, as to whether inputs were turned into outputs (e.g., whether food aid was delivered or wells sunk). Seventy-three per cent were rated as satisfactory or better in assessment of sustainability/connectedness. This impressive rating suggests that evaluators are familiar with these concepts and for the most part have the ability to assess these areas, even though a significant minority of reports are under-performing. Only one report in the case of effectiveness, and two in the case of sustainability/connectedness, did not address these issues.

Problems with assessment of efficiency related mainly to a majority of reports not considering whether the intervention might have taken a less costly route to achieve its objectives – for example, whether different forms of procurement or logistics might have been more cost effective. However, as noted in Chapter 3, a minority of reports did cover areas such as differential costs between international and national staff, and local and international procurement.

The main problems with assessment of impact – where 65 per cent of reports were assessed as unsatisfactory or poor – was the inability of evaluators to look beyond the specific outputs of the intervention to the wider horizon and to examine any unintended consequences, whether positive or negative. One of the areas most often missed was consideration of how interventions were likely to affect socioeconomic relations over the longer term, including gender relations. However, this rating should be read in the context of general difficulties with the assessment of impact in the evaluation field, and it is usually acknowledged as one of the more difficult areas to evaluate.

'Coherence' is the least understood of the OECD/DAC criteria, and is often confused with 'coordination'. This is linked to the failure of evaluators to consider agency policy (see Section M2.4). Indeed the idea of considering whether a number of agencies' policies and strategic directions are similar was beyond the scope of most evaluations – perhaps because many looked at single agency interventions, and no good practice was identified.[10]

Findings, Conclusions and Recommendations M2.7

Overview

Evaluations were weak in terms of making proactive efforts to disseminate report findings, in particular to primary stakeholders, as well as in attempting to ensure that

Table M10 Quality of Findings, Conclusions and Recommendations

Area of Enquiry	Rating	% of Reports Attaining Rating (rounded)
5.4i		
Quality of the sharing of the evaluation findings. a. Preliminary findings should be discussed with key stakeholders, including primary stakeholders, as the evaluation progresses.	Good	0
	Satisfactory	3
	Unsatisfactory	38
	Poor	59
5.4i		
Quality of the sharing of the evaluation findings. b. The draft evaluation report should be shared with key stakeholders, and feedback integrated into the final report or included as an Annex.	Good	0
	Satisfactory	3
	Unsatisfactory	13
	Poor	84
5.4ii		
Quality of conclusions arising from findings.	Good	17
	Satisfactory	45
	Unsatisfactory	28
	Poor	10
5.4iii		
Quality (including feasibility) of recommendations. a. Recommendations should respond to the main conclusions; reflect consultation with all key stakeholders; and understanding of the commissioning organisation; and potential constraints to follow-up. They should be clear, relevant and implementable with each ideally accompanied by implementing options.	Good	22
	Satisfactory	61
	Unsatisfactory	33
	Poor	3
5.4iii		
Quality (including feasibility) of recommendations. b. The evaluation report should suggest a prioritisation (e.g., into macro or structural, micro or easily achievable) and timeframe for follow-up and suggest where responsibility should lie if this is not indicated in the TOR.	Good	0
	Satisfactory	6
	Unsatisfactory	63
	Poor	31

recommendations were followed up and lessons learnt. They thus undermine the whole purpose of evaluation. When accepting evaluation contracts evaluators have a responsibility to attempt to improve agency performance. In general, more time needs to be spent working with agencies to establish what kinds of recommendations are feasible, what recommendations should be prioritised, and who should be responsible for follow-up. Recommendations are too often tacked on in long lists at the end of reports; worse, they are dispersed throughout the report and not included in the Executive Summary.

Sharing of findings (5.4ia & b)

A significant majority of reports include brief details as to feedback mechanisms, usually through end-of-mission meetings with agency country staff, workshops in-country, and debriefings in the HQ of the evaluated agency. There were also some, although fewer, details on circulation and feedback of report drafts. Both mechanisms are now often built into evaluations as a matter of course, but the reports are largely silent on how interaction during debriefings, and comments on drafts, affected the final conclusions and recommendations – even though it is well known that there is usually a period of bargaining between evaluators and commissioning agencies between draft and final versions of a report, particularly concerning phrasing and inclusion of unwelcome findings.

In this year's QP a requirement was added that evaluators needed to share preliminary findings with primary stakeholders (5.4ia) in order to be considered satisfactory. Only one evaluation managed this – the evaluation of Oxfam's intervention in Burundi (March 2002) where the evaluators presented evaluation results to stakeholders, including community hygiene and water committee members, local and national government representatives, and donor and INGO staff.

Part of the reason for the failure to take evaluation findings back to primary stakeholders for discussion and verification probably stems from the format of evaluation missions, which usually start and end in the national capital. There is also usually a major gulf between primary stakeholders and national capital-based staff of NGOs, donors and governments, which makes inviting primary stakeholders to national capital feedback meetings an option rarely considered. This means that what is generally considered good evaluation practice - verifying results with stakeholders - does not take place in EHA.

Quality of conclusions (5.4ii)

Sixty-two per cent of reports rate as satisfactory or good in terms of conclusions flowing logically from, and reflecting, the report's central findings. The assessors mainly considered whether there was a clear connection between findings and conclusions and did not take into account the basis of findings, which is covered mainly under Section M2.2 above. If the latter area had been considered then the rating on conclusions would have been considerably lower.

Quality of recommendations (5.4iiia & b)

This year two separate aspects were included in the attempt to differentiate particular strengths and weaknesses: clarity and quality. Writing recommendations is an art. It is perhaps even the most important part of evaluation practice and not something evaluators can be expected to do without training or guidance.[11] While some of the failure of uptake of recommendations is due to political factors within and between agencies – itself something evaluators should be aware of when writing recommendations – this also partly results because recommendations are inadequately crafted.

In 5.4iiia, the assessors mainly focused on whether recommendations were clearly written, relevant, and responded to the main conclusions. While this section of the QP also included a requirement that recommendations be implementable and demonstrate an understanding of the commissioning organisation, it was not possible for the assessors to judge this accurately given the wide range of countries and agencies involved.

Eighty-four per cent of reports were rated as satisfactory or better in terms of this first area, the highest rating of any QP Area of Enquiry. Recommendations tended to be clearly phrased and followed on from conclusions

However, in relation to 5.4iiib, no reports met the requirement of producing recommendations that were: a. prioritised; b. included a timeframe for follow-up; and c. suggested where responsibility for follow-up should lie. Many reports included long lists of recommendations, sometimes stretching to several pages and sometimes dispersed unhelpfully through the report.

In light of this, evaluators should consider taking a more proactive approach to the writing of recommendations:

- noting four or five recommendations they see as central in the Executive Summary section;
- providing a suggested timeframe for each of these recommendations;
- naming specific agency positions (e.g., project manager) responsible for follow-up; if that is not possible, a department or unit.

Quality of Report Coverage, Legibility and Accessibility M2.8

Table M11 Report Coverage, Legibility and Accessibility

Area of Enquiry	Rating	% of Reports Attaining Rating (rounded)
6.i		
Quality of the coverage of the evaluation report.	Good	23
	Satisfactory	35
	Unsatisfactory	42
	Poor	0
6.ii		
Quality of the format of the report.	Good	15
	Satisfactory	19
	Unsatisfactory	7
	Poor	59
6.iii		
Accessibility of the report.	Good	21
	Satisfactory	47
	Unsatisfactory	29
	Poor	3
6.iv		
Quality of the executive summary.	Good	21
	Satisfactory	46
	Unsatisfactory	21
	Poor	12

Quality of report coverage, legibility and accessibility (6.i & ii)

Note: the analysis of these two areas of enquiry only involves the 28 reports that included a TOR as coverage and format were rated against the requirements of the TOR.

With reference to the first area, evaluation reports were required to cover adequately all areas specified in the TOR in addition to any further factors likely to effect the performance of the intervention. The fact that in 58 per cent of cases where TOR were present there were areas missed by evaluators suggests that commissioning agencies are not using the TOR as a means of holding evaluators accountable to their agreement with the agency. Across the areas covered in the TOR, there was no one area that stood out as consistently missed by evaluators. However, five reports did not cover one or more of the DAC criteria, and three did not cover gender equality issues, all as required in the TOR.

In terms of report format, a majority of reports were rated as unsatisfactory or worse as the TOR did not provide a template format to follow (although in a number of cases, including the WFP reports, a required format is mentioned in the TOR but not included in the version received by ALNAP). Commissioning agencies not including a required format miss an important opportunity to provide evaluators with guidelines as to their priorities, as well as to promote greater attention to areas generally missed in EHA such as protection and gender equality. An exception were the ECHO reports which were found to be generally strong in both setting a required format and ensuring that this was followed.

Accessibility of the report and Executive Summary (6.iii & iv)

Reports were generally well written in clear English or French, although more reports could have included visual aids such as maps, tables and diagrams. A number of reports included long stretches of unbroken text, trying for even the most patient and interested reader. The quality of Executive Summaries was satisfactory or better in 67 per cent of cases, and removing those four reports where no Executive Summary was included (a major oversight), performance in this area can be considered adequate. For the 21 per cent of reports rated as unsatisfactory, the main issue was failure to include all key report recommendations in the Executive Summary. There were no significant differences between ECHO, the UN and NGOs in these areas, except in the case of ECHO reports in English, which tended to lack clarity.

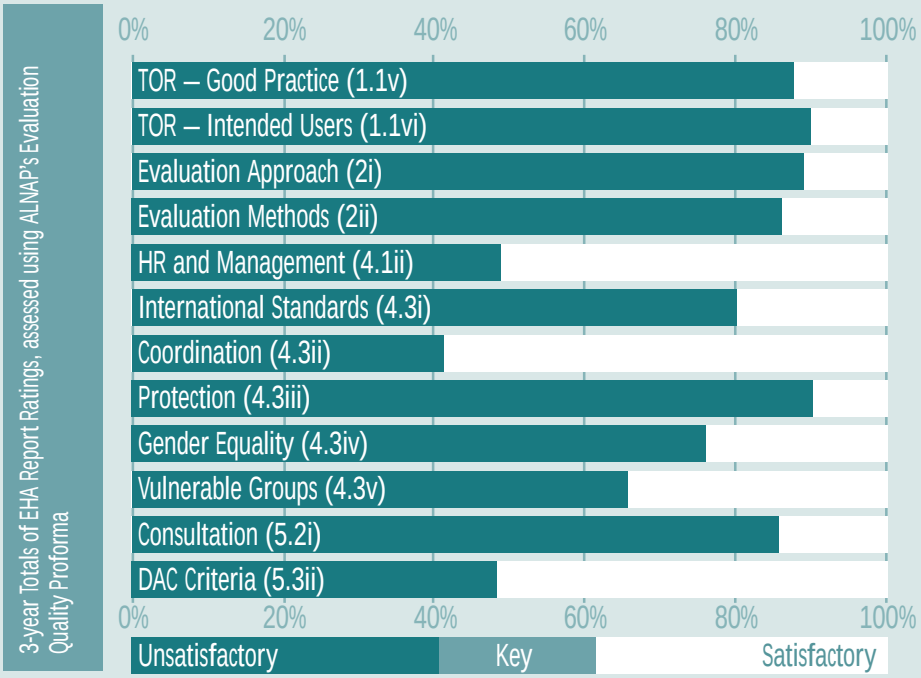
Year-on-year Analysis of 127 Evaluations of Humanitarian Action (2000-2002)

M3

Box M3 Year-on-year Analysis of 127 Evaluations of Humanitarian Action (2000–2002)

This year we report on some comparative areas that are key to successful evaluation over the period of the three Annual Reviews, based on the 127 reports assessed against the QP for 2000–2002. Methodological details related to this comparative analysis can be found in Section M1.6 above. The criteria against which reports were rated can be seen in the Guidance Notes section of the QP at the end of this section.

Twelve QP Areas of Enquiry are compared, as set out in the Figure below where aggregate results for the three year period are presented.



Box M3 Year-on-year Analysis of 127 Evaluations of Humanitarian Action (2000–2002) *continued*

TOR	Satisfactory %	Unsatisfactory %
Quality of TOR statement on expectation of good practice in approach and method	11	89
Quality of TOR statement on intended use and users of evaluation outputs	8	92

Evaluation reports were found to be weak in both of these areas. In general, reports did not specify adequately the key methodological tools that evaluators should use. It was also rare for TOR to outline clearly the intended use of evaluation reports; failure to do this adds to the likelihood that the findings of these reports will not be fully used.

Delineation of Methodology	Satisfactory %	Unsatisfactory %
Appropriateness of the overall evaluation approach	10	90
Appropriateness of the evaluation methods selected	12	88

In terms of outlining, explaining and providing a rationale for the evaluation approach, performance was generally unsatisfactory, in particular in 2002 and 2003. EHA is atheoretical and as such derives little direction from wider evaluation thinking. For example, the debate over the relative emphasis to be placed on lesson learning and accountability in EHA has also been taking place in the wider evaluation field, but EHA practitioners have made few linkages. This is not to suggest that every evaluator needs to become a specialist in evaluation theory. Far from it. But commissioning agencies and evaluators do need to have a broad understanding of the strengths and weaknesses of different evaluation approaches so as to avoid common pitfalls and make EHA as rigorous as possible.

A small minority of reports achieved good practice in terms of delineating the methodology that was to be used. However, most reports note only basic details of the methodology, which in turn undermines the credibility of their findings.

Box M3 Year-on-year Analysis of 127 Evaluations of Humanitarian Action (2000–2002) *continued*

	Satisfactory %	Unsatisfactory %
Quality of the Evaluation of Agency's Management and Human Resource Practices	51	49

This is a strength in EHA, with over 50 per cent of evaluations rating as satisfactory or better each year. Evaluators have consistently examined issues such as staff turnover, HQ-field communication, and security. However this can be considered both a strength and a weakness because the focus on institutional issues may detract from other areas, such as consultation with primary stakeholders or international standards.

Cross-cutting Themes	Satisfactory %	Unsatisfactory %
Evaluation of use of international standards	20	80
Evaluation of co-ordination	58	42
Evaluation of protection	10	90
Evaluation of gender equality	26	73
Evaluation of consideration to vulnerable/ marginalised	36	64

The cross-cutting theme that consistently scored well was coordination, which is related to the ability of evaluators to cover institutional factors. In the other four theme areas reports performed consistently poorly except in the case of consideration to the vulnerable and marginalised where performance was somewhat better. The link between international standards, protection and gender equality is that they deal with rights-based issues that are often controversial; these are the issues that are most often left out of evaluation TOR and with which evaluators appear to have the least skills. Protection is particularly poorly covered, with 92 per cent of the reports in 2002 and 79 per cent of reports in 2001 assessed as unsatisfactory or poor. The Red Cross/Red Crescent Code of Conduct and the Sphere standards are also not generally used.

This is a central gap in EHA, which is clearly a long way away from integrating a rights-based approach into a wider evaluative process. ALNAP can play an important role in terms of getting this issue on the agenda of commissioning agencies and evaluators.

Box M3 Year-on-year Analysis of 127 Evaluations of Humanitarian Action (2000–2002) *continued*

	Satisfactory %	Unsatisfactory %
Quality of Consultation with and Participation by Primary Stakeholders	13	87

This is a further area of weakness. Despite some good practice, EHA could rightfully be accused of systematically ignoring the views and perspectives of primary stakeholders in favour of those of institutional actors, particularly agency staff. This undermines its credibility and continues in the vein of treating primary stakeholders as passive recipients of aid rather than active participants in their own recovery. This agency-centric perspective will only change if commissioning agencies insist on adequate primary stakeholder consultation and participation. The constraints to this, particularly security issues, should not of course be underestimated. But an equally important constraint would appear to be the structure of evaluation missions which are usually short forays by foreign-based evaluators, with a focus on national capitals.

	Satisfactory %	Unsatisfactory %
Application of the DAC Criteria	50	50

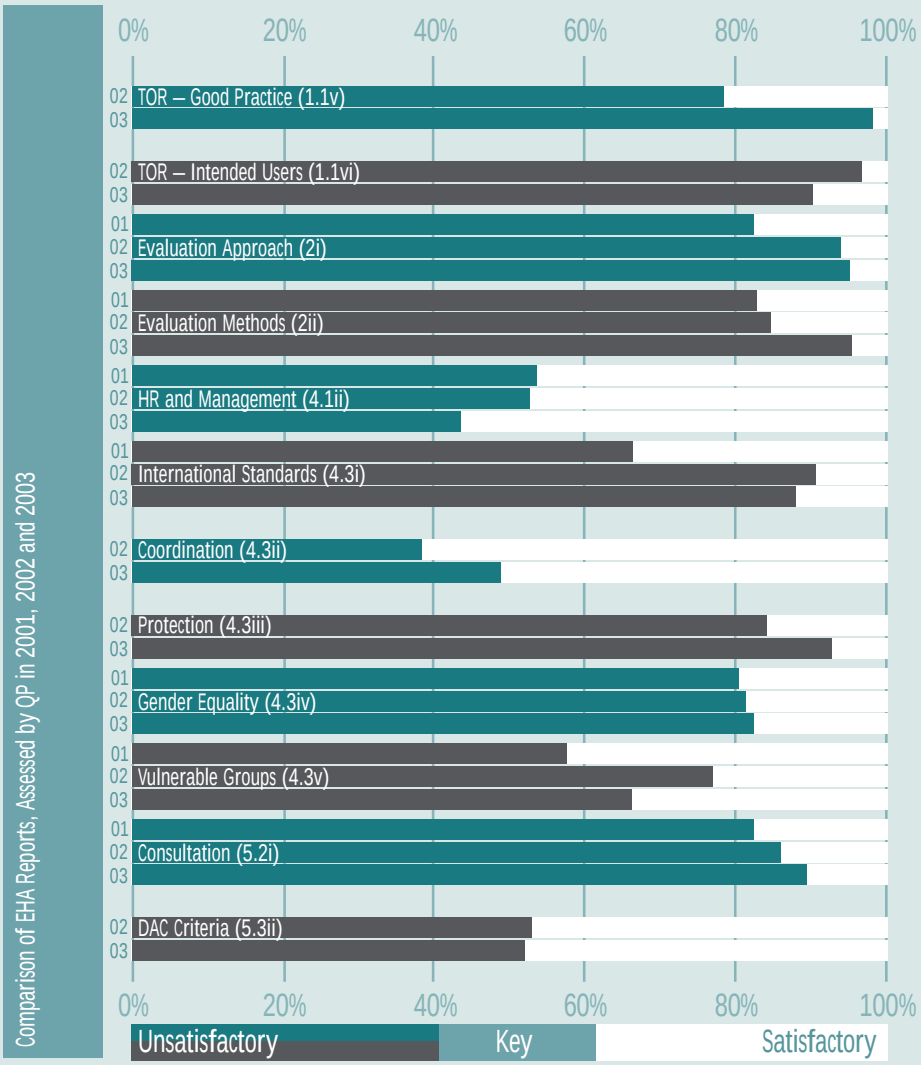
This analysis is based on reports assessed for this and last year only, as the QP for 2001 did not use comparable phrasing. Results for this year have been aggregated across the seven criteria. Much EHA is organised around the DAC criteria, as reflected in most evaluation TOR. Application of the DAC criteria is one of the stronger areas of EHA, with the third highest rating of the 12 areas covered in this Box. Overall it is possible to conclude that evaluators have had reasonable success with their application and that they have become EHA's central evaluative tool.

Year-on-Year Improvement?

As several of the areas of enquiry cover only this and last year, and because of changes in the Proforma over time, it is difficult to come to definitive conclusions concerning year-on-year improvement in EHA. The areas where there may have been some improvement are in the evaluation of management and human

Box M3 Year-on-year Analysis of 127 Evaluations of Humanitarian Action (2000–2002) *continued*

resources as well as the mainstreaming of DAC criteria. In this latter area, however, only some of the criteria are being consistently used (see Section M2.6).



Given the emphasis on results-based planning in many agencies, one would expect to see ongoing improvement in EHA. Over the next two years it will be important to assess progress in key evaluation areas such as attention to international standards, gender equality, and consultation with and participation of primary stakeholders. Some suggestions on standards and target-setting for agencies are included in the conclusions to this meta-evaluation (Section M3).

Interagency Differences?

Finally, were there any marked differences between the UN system, ECHO and NGOs as far as evaluation quality is concerned? All actors have their strengths and weaknesses, some of which can be highlighted as follows:

- ECHO and NGOs did relatively well in assessment of coordination, management and human resources and attention to the vulnerable, but relatively poorly in detailing evaluation processes and paying attention to adherence to international standards, protection and gender equality.

- Overall, UN agencies performed best in 10 of the 12 QP areas considered in this Box (only six areas are covered), often by a considerable margin. This relates in particular to strong evaluation performance by WFP and UNHCR. Even so, UN agencies failed to achieve a 50 per cent satisfactory rate in six of the areas considered.

Conclusions M4

This year's assessment of 39 reports revealed some improvement in evaluation performance, but also highlighted ongoing weaknesses. Good practice, this year in the case of WFP and the DEC, illustrates what is possible given resources, capacity and mindset. A common theme this year has been the need to understand and measure changes in social processes more thoroughly – in particular, power relations, gender relations and indigenous coping strategies. Commissioning agencies and evaluators should reflect on whether their evaluation's consideration of these areas is adequate.

Some of the key areas commissioning agencies and evaluators should pay attention to over the next year are as follows:

Evaluation Focus

- Ensure that protection issues and reference to international standards are included in TOR, where relevant.
- Bring to the evaluator's attention relevant agency policies, including the gender equality policy if it exists, and ensure that the need to evaluate against agency policy is clearly set out in the TOR.
- Adequately evaluate primary stakeholder participation and consultation.
- In the context section of the report, note the past involvement of the agency in the affected area, any partnerships that have built up, and how these affected the intervention.
- Pay particular attention to the DAC criteria which may be less well covered in evaluations, in particular impact, efficiency, and coherence.
- Ensure that data in reports is disaggregated by socioeconomic status, ethnicity and sex.

Evaluation Process

- Look for innovative ways to disseminate report findings, for example, through thematic summaries or key sheets. Follow-up informally with colleagues to see if recommendations have been followed. If they have not, analyse why.
- Promote primary stakeholder consultation and participation, and ensure that there is a requirement to do this in the TOR.
- Ensure that the methods used provide a credible basis for conclusions and that the description of the method fully reflects what the evaluation team has done – in particular in relation to consultation and participation of primary stakeholders, and the nature of participation of other key stakeholders.
- Note how confidentiality and dignity of respondents is ensured.
- Establish or build on contacts with evaluators from affected countries and consider including them in evaluation teams or making it a requirement in

tenders that at least one person from the affected country be included on the team.

- Publicise the cost of the evaluation in the TOR so as to allow a comparative analysis of evaluation costs and to ensure that adequate resources are being allocated for evaluation purposes.

Systemic problems in the quality of EHA have been identified by the three year comparative analysis in Box M3. This suggests that ALNAP member agencies should, among other initiatives, consider developing a set of standards for improving their evaluation practice in some of the weaker areas of EHA. While potentially controversial, it is this author's view that it is unlikely that there will be a significant improvement in evaluation practice over the next few years based on capacity development alone. This is because unsatisfactory practice is resulting not only from lack of capacity, but also because commissioning agencies are not consistently enforcing good practice requirements.

Agencies are in many cases already committed – through their policies and evaluation guidance – to covering adequately a number of the weaker areas in EHA, such as consultation with and participation of stakeholders, gender equality, use of good practice in methodology, and international standards. So target-setting to meet EHA standards would also be an accountability mechanism to ensure that agencies fulfil their commitments. Standards could be adapted from the QP, as have the lists above; if the idea of target-setting against these standards is adopted in principle the specifics would need to be discussed in the ALNAP forum by all ALNAP Full Members.