

**THE 'MEFF' METHODOLOGY:  
A REVIEW OF DFID'S MULTILATERAL EFFECTIVENESS FRAMEWORK**

**CONTENTS**

	ACRONYMS	2
	EXECUTIVE SUMMARY	3
1	INTRODUCTION	5
2	WHY DID WE DEVELOP THE MEFF?	5
3	IN-HOUSE OR CONTRACT OUT?	6
4	HOW TO ASSESS EFFECTIVENESS?	7
5	DESIGN OF THE MEFF	10
6	THE MEFF INSTRUMENTS	11
7	CONSULTATION PROCESS	13
8	QUALITY ASSURANCE	13
9	EVALUATION OF THE METHODOLOGY	14
10	TRANSACTIONS COSTS	23
11	VIEWS OF THE MULTILATERALS	28
12	RISKS AND SUSTAINABILITY	31
	Annexes	
1	MEFF figures	34
2	Rules of aggregation	39
3	Agency feedback	40
4	References	41

Alison Scott  
International Division Advisory Department  
Department for International development (DFID)  
March 16 2005

## Acronyms

AfDB	African Development Bank
AfDF	African Development Fund
AsDB	Asian Development Bank
AsDF	Asian Development Fund
CDB	Caribbean Development Bank
CMPS	Centre for Management and Policy Studies
EBRD	European Bank for Reconstruction and Development
EC	European Commission
EDF	European Development Fund
EIB	European Investment Bank
DAC	Development Assistance Committee (OECD)
FAO	Food and Agricultural Organisation
Habitat	United Nations Human Settlements Programme
IADB	Inter-American Development Bank
HR	Human Resources
ICRC	International Committee of the Red Cross
IDA	International Development Association
ID	International Division
IDAD	International Division Advisory Department
IFAD	International Fund for Agricultural Development
IFRC	International Federation of the Red Cross
ILO	International Labour Organisation
IPPC	Integrated Pollution Prevention and Control
IS	Institutional Strategy
ISP	Institutional Strategy Paper
MDB	Multilateral Development Bank
MDGs	Millennium Development Goals
MEG	Multilateral Effectiveness Group
M&E	Monitoring and Evaluation
MEFF	Multilateral Effectiveness Framework
MOPAN	Multilateral Organisations Performance Assessment Network
NAO	National Audit Office
NGOs	Non-Governmental Organisations
OCHA	Office for the Coordination of Humanitarian Affairs
OHCHR	Office of the High Commissioner for Human Rights
PGRFA	Plant Genetic Resources for Food and Agriculture
PRS	Poverty Reduction Strategy
PRSPs	Poverty Reduction Strategy Papers
PSA	Public Service Agreement
QA	Quality Assurance
RBM	Results Based Management
RMC	Regional Member Countries
SPA	Special Partnership with Africa
SWAPs	Sector Wide Approaches
UNDP	United Nations Development Programme
UNESCO	United Nations Education, Scientific and Cultural Organisation
UNFPA	United Nations Population Fund
UNHCR	United Nations High Commissioner for Refugees
UNICEF	United Nations Children's Fund
UNIDO	United Nations Industrial Development Organisation
UNIFEM	United Nations Development Fund for Women
UN	United Nations
WFP	World Food Programme
WHO	World Health Organisation

## EXECUTIVE SUMMARY

- i. During 2003-04, DFID set up a multilateral effectiveness framework (known as the MEFF) for assessing the organisational effectiveness of the multilaterals it supports centrally. The system was developed internally and DFID staff conducted the assessments. Twenty-three organisations have been assessed.
- ii. The purpose of the MEFF is to provide information for DFID's reporting on its Public Service Agreements (PSA), its Institutional Strategies (ISs) with multilaterals, and its financing decisions. An overview paper sets out the main results and how they will be used in the future. The focus of this paper is to describe and assess the methodology.
- iii. The MEFF focuses on **organisational effectiveness**, using a results-based management (RBM) approach. It tells us about the enabling conditions that are necessary for the achievement of results on the ground, although it does not tell us about how well these organisational systems are implemented, or what results are actually achieved. It is only one part of the picture of an organisation's effectiveness. Section 4 outlines the analytical reasons for this approach.
- iv. The MEFF focuses on eight organisational systems and looks at them in terms of their focus on internal performance, country level results and partnerships. The main assessment instruments are a checklist with 72 questions, a scorecard and a summary report. Sections 5-8 describe the basic design and implementation processes.
- v. Sections 9-12 evaluate the approach, design and application of the assessment instruments, the transactions costs involved, the views of the assessed multilaterals and the sustainability of the exercise. The findings are that:
  - The overall approach is valid as it does identify important differences amongst multilaterals that affect their performance.
  - The checklist questions were well designed on the whole, and the modification for humanitarian agencies has worked well.
  - There are some limits to the generic applicability of parts of the framework, particularly for coordination agencies and agencies that do not have autonomous governance arrangements. The relevance of some issues, such as PRSPs, country level programming, and the coverage of global standards work has been questioned by some agencies. However, the overall number of 'blank scores' is very small.

- The answers have met our standards of objectivity and accuracy, and have been well evidenced.
  - The integrity of the scores is high; initially there were a small number of questionable scores but these have been adjusted in the interests of inter-agency consistency.
  - The system for aggregating scores had problems initially, but these were mostly ironed out during the quality assurance process.
  - The summary reports have received less attention so far, and will need further work to identify indicators for monitoring progress in three areas over the next two years. A full revision of the MEFF baseline is not expected for three years.
  - The cost in terms of staff time and travel has been low, relative to an external consultancy exercise. However, as the exercise was undertaken on a part-time basis with no additional staff resources, the main cost has been delay in completing the work.
  - There have been considerable benefits to conducting the exercise in-house, in terms of staff ownership of the results and an immense internal learning about the multilaterals, results-based management and effectiveness issues generally.
  - We adopted a transparent and consultative approach with the assessed agencies, which was greatly appreciated by them. However, the quality of the MEFF has relied greatly on their direct participation – which has implied higher transactions costs for them than we had intended.
  - Feedback from the agencies indicates extensive support for the MEFF in terms of the organisational approach, the three perspectives and (on the whole) the indicators used. However, they have concerns about the potential implications of a proliferation of similar initiatives and urge us to promote bilateral coordination on such assessments. We will be taking forward this recommendation with our bilateral partners.
  - The main risk to the sustainability of the MEFF comes from staff turnover, insufficient resourcing of overhead costs and lack of awareness of the MEFF in other parts of DFID. Internal training and dissemination will be required.
- vi. Next steps include developing guidance on how the MEFF should be integrated with the ISs and identifying indicators for the three MEFF monitoring areas. A full revision of the baseline will take place in 3-4 years time, unless an agency requests one earlier.

## 1. INTRODUCTION

1.1 In 2003-04, DFID's International Division (ID) established a Multilateral Effectiveness Framework (known as the MEFF) for assessing the organisational effectiveness of the multilaterals that it funds centrally. It was developed over 21 months, and has been used to assess twenty-three organisations.<sup>1</sup>

1.2 The main objectives of the MEFF are to:

- Provide an information and monitoring system that would support DFID's reporting on its Public Service Agreement (PSA) objectives
- Provide inputs to DFID's corporate engagement with multilaterals via Institutional Strategies (ISs)
- Provide inputs to future financing decisions

1.3 This paper reviews the approach and methodology of the MEFF, with a view to documenting and assessing the decisions that were taken and the process through which it was implemented. A companion overview paper summarises the main results and how they will be used by DFID (Scott 2005).

## 2. WHY DID WE DEVELOP THE MEFF?

2.1 DFID's concern with assessing multilateral effectiveness reflects an increased focus on performance and effectiveness in international aid agencies, and more generally in the private sector and in government. Most organisations are now adopting a results-based approach to performance, broadening out from narrower concerns with internal efficiency and value for money. They are developing results-based management (RBM) systems, which set objectives and targets for their internal business processes and use them for improved delivery, management and accountability.

2.2 The international development community is addressing these issues as part of the common effort to achieve the internationally agreed Millennium Development Goals (MDGs). Aid agencies are examining how they can contribute to better outcomes in developing countries, and what systems they can put in place to make this contribution more effective. In the UK, there is increasing pressure from the Treasury, Parliament and civil society for DFID to account for its expenditures in terms of results in developing countries.

2.3 DFID will use the information on multilateral effectiveness as an input to its reporting on PSA targets, its Institutional Strategies, and its financing decisions. DFID first introduced Public Service Agreement (PSA) targets in 1999, and has revised them twice since then. In 2002, a new PSA objective was introduced for the International Division, 'to increase the impact of key multilateral agencies in reducing poverty and effective response to conflict and humanitarian crises'. One of the targets for this PSA was to improve the effectiveness of the international system as demonstrated, *inter alia*, by

---

<sup>1</sup> The UNESCO MEFF was completed recently and the results are still provisional

working to improve the institutional effectiveness of 12 multilaterals.<sup>2</sup> ID would need to report to the Treasury on progress with this target in the Autumn Performance Report and the Departmental Report.

2.4 Since 1999, DFID's corporate engagement with multilaterals has been set out in Institutional Strategies for each agency that it finances.<sup>3</sup> In 2002, the National Audit Office (NAO) commented that the performance management of these strategies should be tightened up (NAO 2002:29-30). Accordingly, new guidance was issued in May 2003, with an increased focus on effectiveness issues. So the assessment and monitoring of effectiveness would extend to all 26 ISs,<sup>4</sup> not just the 12 PSA agencies.

2.5 In 2002, multilaterals received almost half of DFID's budget from central funding by the International Division, and substantial further amounts from country programmes.<sup>5</sup> The Finance and International Divisions were interested in developing better criteria for financial decisions, as regards the split between bilateral and multilateral expenditures and expenditures for different multilaterals. This information was also required as part of the discussions on the 2004 Departmental Spending Review.

2.6 These external and internal pressures for improved management of our engagement with the multilaterals required the establishment of an evidence base on multilateral effectiveness, from which to track trends over time. Figure 1, Annex 1 shows how this evidence base would feed into Institutional Strategies, PSA reporting, financing and other decisions.

2.7 There was no precedent for this type of work. Although individual evaluations of some multilaterals are carried out on a periodic, once-off basis, there have been no systematic, comparative, across-the-board assessments. More fundamentally, there is little international consensus on *how* such an assessment should be carried out. There is little agreement on how to define multilateral effectiveness, far less on how to measure it.

### **3. IN-HOUSE OR CONTRACT OUT?**

3.1 We considered that it would be difficult to contract this work out to consultants, as we were not entirely sure what we wanted. The right mix of consultancy skills (knowledge of RBM, development and organisational effectiveness, familiarity with DFID's agenda and with the international/multilateral aid system) appeared to be scarce. There was no readily available off-the-shelf methodology that could be accessed via consultants.

---

<sup>2</sup> This is known as Service Delivery Agreement VII. The 12 agencies were: The World Bank, AfDB, AsDB, EBRD, ICRC, UNDP, UNICEF, UNFPA, WHO, FAO, UNHCR, and UNESCO. They were selected because they receive over £10 million each, annually from DFID.

<sup>3</sup> The first generation Institutional Strategies were referred to as Institutional Strategy Papers (ISPs); subsequently the 'Paper' was dropped and they are now referred to as ISs.

<sup>4</sup> In addition to the 12 agencies, DFID has ISs with IADB, CDB, Habitat, UNIFEM, UNIDO, IFAD, WFP, IFRC, OCHA, UNAIDS, OHCHR, the EC and EIB and the Commonwealth Secretariat.

<sup>5</sup> DFID Departmental Report 2003, page 128.

3.2 On the other hand, there was a strong case for undertaking the work in-house. Many of the ISPs developed in 1999-2000 focused on supporting organisational change in the multilaterals in order to improve their effectiveness.<sup>6</sup> The desk officers who managed these Strategies would be the main users of the MEFF information for future IS management and PSA reporting. Many staff were involved in discussions with other shareholders on the Boards and Governing Bodies about performance and effectiveness.<sup>7</sup> The IS desk officers had some knowledge of the organisations concerned, and it seemed at the outset that it would be easy for them to find the information for the assessments. Many of them had already been involved in an earlier in-house attempt to measure multilateral effectiveness (Dyer et al. 2003) and had participated in a parallel work stream on the Vision for the International System (Turner et al. 2003).

3.3 Accordingly, the decision was taken to develop the MEFF internally. It was led by the International Division Advisory Department (IDAD), with the participation of the IS desk officers, individually and through an internal, cross-departmental working group, the Multilateral Effectiveness Group (MEG). Some 35 staff were involved in the MEFF over the whole period (19 IS desk officers at any one time) – an unprecedented scale of internal participation in a research type activity of this sort. The work proceeded on an experimental, learning-by-doing basis. It grew organically; we set out on a journey with a vision of the destination, and worked out how to get there on the road.

3.4 The MEFF work evolved as follows:

- Initial concept note developed by IDAD (March 2003)
- Analytical work developing the approach (April-August)
- MEG set up (May)
- Methodology developed and piloted (September-December)
- Consultations with some multilaterals - World Bank, ILO, UNDP, UNICEF (June, September)
- Scorecard training (October and December)
- Agency presentations (mainly January–March 2004)
- Desk officers implement MEFF assessments in consultation with agencies (January-June)
- Formal quality assurance (July-October)
- Baseline assessments completed for 22 agencies (October 22)
- Results analysis (December)
- Internal Review (January-February 2005)
- External presentations to Whitehall, Bilaterals etc, sporadically throughout this period

#### **4. HOW TO ASSESS AGENCY EFFECTIVENESS?**

---

<sup>6</sup> E.g. UNDP, UNICEF, UNIDO, some humanitarian agencies.

<sup>7</sup> For example, with IDA Deputies at the World Bank, the WEOG (Western Europe and Others Group) at the UN, the Utstein Group of North European donors etc.

4.1 The debate about effectiveness is bedevilled by confusion over concepts and definitions. First, we need to distinguish between *aid effectiveness* and *agency effectiveness*. The former refers to the aggregate impact of aid (e.g. financial flows, technical assistance) without looking at the contribution of particular agencies, while the latter looks at the role of these agencies. Second, we need to define agency effectiveness – but this is not easy. There are many ways to assess the effectiveness of an agency: what impact does it have on its intended beneficiaries? Does it produce high quality outputs? Does it have good policies? How well does it behave as a partner? Is it well managed? Does it have well functioning organisational systems? These different questions require different types of evidence and assessment methods.

### **Difficulties in assessing on the basis of results**

4.2 Intuitively, one might expect to be able to judge agency effectiveness in terms of results on the ground and to attribute these results to the actions of particular aid agencies in the same way as exam results may be attributed to schools. However, it is difficult to assess agencies in this way for three reasons:

- There is very little information available on results;
- There are problems in attributing results to the actions of a single aid agency
- It is difficult to compare different types of results (e.g. inoculations, enrolments, policy advice, capacity building).

4.3 Our initial survey of official agency reports revealed that there is very limited reporting on results by multilaterals.<sup>8</sup> Most of them report on their activities and inputs (especially financial), rather than outcomes. However, even if this information were forthcoming, it would be difficult to attribute the outcomes to individual agencies. This is because the relationship between inputs and outcomes is much less direct than between schools and pupils. These days, most aid is delivered through intermediaries rather than direct to the poor (see figure 2, Annex 1). Therefore outcomes are the product of multiple influences (by governments, donors, civil society, etc.), as well as external factors such as climate, disease or currency fluctuations. These problems were recognised by NAO in its assessment of DFID's performance management system, where only 9% of its 2001-04 performance measures were considered to be attributable to DFID:

'With the majority of measures it is not possible to determine the extent to which any achievement is a result of DFID's effort, because of the numerous other factors and organisations in development work.' (NAO 2002:26).

---

<sup>8</sup> In preparation for this work, IDAD reviewed the reports of the 12 PSA agencies, to see to what extent information on agency results was available. The study concluded that most reports were input and activity oriented, and information on outcomes and impacts was rare.

4.4 The problem of attribution is now widely recognised in the development community (Meier 2003, World Bank 2002, Farquhar 2000, Binnendijk 1999). But that does not mean that the effort to trace results should be abandoned. Aid agencies need to identify the *intermediate* outcomes for which they are more directly responsible, and show how these outcomes contribute to longer and more complex causal chains (along with the contribution of other stakeholders). However, this practice is relatively new, so there is little evidence yet on intermediate outcomes (OECD/DAC 2005a).

4.5 There is also the problem of non-comparability of agency results arising from differing mandates and functions. Comparing the results achieved by Development Banks with UN specialised agencies, humanitarian agencies, and international charities would be like comparing outcomes from the Departments of Health, Work and Pensions, the Bank of England and Shelter.

### The RBM approach

4.6 In the absence of information on results, the next best option – as a partial measure – is to focus on the *enabling conditions that are necessary for the achievement of results*, i.e. having in place and implementing organisational systems that focus on results. This approach is based on results-based-management (RBM) theory, which assumes that an effective organisation is one that incorporates a results focus into all its business processes (e.g. strategic planning, resource allocation, HR management, M&E) and uses the information on results to continually improve its performance. Thus there is an assumed link between these enabling conditions and actual results.<sup>9</sup> The focus on organisational systems is referred to as *organisational effectiveness*. NAO calls it '*organisational capability*':

'Other public sector organisations, have, however, started to adopt packages of performance indicators which balance results indicators with indicators of **organisational capability – which can help assess prospects for future results** (emphasis added). An approach of this sort could have particular value for an organisation such as DFID where results are distanced in time or certainty from activity' (NAO 2002:53)

4.7 An additional advantage of an RBM approach is that it provides the basis for a generic, comparative assessment across different agencies. It assumes that all aid agencies have common ways of organising themselves, regardless of mandate and function, in order to maximize the potential for results (e.g. strategic focus, efficient resource management, good quality control, monitoring and evaluation, etc.).

4.8 DFID decided to adopt this approach because it would provide the basis for a comparison of relative effectiveness. It was also consistent with

---

<sup>9</sup> An example of this link is the improved project quality of World Bank projects after they set up the Quality Assurance Group.

DFID's emphasis over the last five or so years, on the importance of organisational reforms amongst multilaterals. The MEFF would focus on whether results-oriented systems were in place, as a minimum requirement for the achievement of results. However, we did not have the resources to look at how well these systems were implemented. That information would have to be acquired later, and preferably produced by the agencies themselves through their own internal review and reporting processes. Therefore the MEFF offers a partial measure of effectiveness, and should ideally be supplemented with information about quality of implementation and results achieved on the ground.

4.9 The analytical approach was developed on the basis of in-house research of the RBM literature, previous consultancies (Flint 2003, Balogun 2003, Bezanson et.al. 2003), previous in-house assessment exercises (Dyer et.al. 2003), and consultations with some multilaterals (ILO, the World Bank, UNDP and UNICEF). An approach paper was drafted as the basis for internal and external discussion during June – August 2003, and this was later amended to incorporate subsequent methodological decisions (Scott 2004).

## 5. DESIGN OF THE MEFF

5.1 The MEFF assesses effectiveness in terms of the results focus of multilateral organisational systems. It looks at **eight systems** in all:

- corporate governance;
- corporate strategy;
- resource management;
- operational management;
- quality assurance;
- staff management;
- monitoring, evaluation and lesson learning;
- reporting of results.

It assesses each of these systems with **three perspectives** in mind:

- their focus on internal performance;
- their focus on country-level results;
- their focus on partnerships.

5.2 We assume that all development or humanitarian agencies have these eight organisational systems, even if they are structured differently in some cases. We also assume that the three perspectives are relevant to all international agencies, i.e. they are concerned about internal performance, have an objective to achieve and track outcomes at country level (even if they are not specifically focused on MDGs or PRSPs), and make some effort to work in coordination with other relevant agencies. The reasoning behind these assumptions was set out in the approach paper (Scott 2004: 6-8).

5.3 In developing the MEFF methodology, we were guided by several key principles:

- *Partnership.* DFID is highly committed to the goal of working with others in order to achieve the MDGs, both at country and international level. We are also keen to influence the multilaterals, through our ISs and other processes, to improve their effectiveness and that of the international system generally. Assessment exercises are potentially confrontational and could risk undermining our relations with the agencies and jeopardise the IS process. Used positively, however, the assessments might promote change. We therefore adopted a transparent, consultative approach with the agencies.
- *Objectivity. Using agencies' own information sources as the evidence.* Previous studies had relied heavily on subjective perceptions about agency effectiveness, which were sometimes ill informed. We would try to reduce subjectivity by emphasising factual information derived from agencies' own information sources (such as planning documents, annual reports, evaluations, etc.). We had neither the intention nor the resources to generate extra investigative work that would risk duplicating information that was already available.
- *Generic criteria.* We would attempt to use criteria and indicators that would be applicable to all agencies, regardless of differences in mandate and function. We draw on generally accepted principles and good practice regarding development effectiveness, but we would try to avoid using any particular institutional model.
- *Comprehensive.* We would not privilege any particular aspect of effectiveness (such as financial efficiency, use of particular aid instruments), but would take a holistic view of the full set of the agency's organisational systems.
- *Recognise strengths while identifying weaknesses.* We would use the results in a way that would maximise the potential for dialogue and change within an agency. We would focus on disaggregated information that directed attention to areas for improvement.

5.4 These principles constrained our methodological choices in a number of ways. It affected the selection of variables, criteria and indicators; the design of the 'traffic light' measurement system; and the consultative process through which the assessments were carried out.

## 6. THE MEFF INSTRUMENTS.

### a) The checklist

6.1 The main assessment instrument is a **checklist**, which is represented as a matrix, with the eight organisational systems listed horizontally and the three perspectives, vertically. This gives a total of 24 cells in the matrix, in each of which there are up to four indicators, expressed as questions. See figure 3, Annex 1.

6.2 The checklist was developed collectively by the Multilateral Effectiveness Group (MEG) during the last six months of 2003, and was piloted internally during this period. The questions/indicators were formulated on the basis of good practice thinking in the RBM and development literatures, as set out in the approach paper.

6.3 A major effort was made to ensure that the questions were relevant to all the assessed organisations. Because of significant differences between humanitarian and development agencies, we developed a separate humanitarian checklist and modified about a fifth of the questions. The checklist criteria were the same but the questions omitted reference to PRSPs, MDGs, alignment to national institutions, and had a slightly different wording on financing, operational management and inter-agency partnerships. This checklist was applied to agencies whose primary function is humanitarian, but not to humanitarian branches of development agencies (such as UNICEF, UNDP, WHO and the EC). WFP is a dual function agency, but was assessed as a humanitarian one because development is only a small part of its work.

6.4 In framing the questions, we tried to avoid evaluative language that would require a subjective judgement on the part of the assessor. The questions are deliberately factual in nature: they ask whether a system or practice is in place or not or is under development. In a few areas, where it was not possible to avoid a judgement, additional factual criteria were suggested to guide the assessment. Guidelines were developed to provide key definitions and rules for completing the assessment.

6.5 The answers were inputted electronically into the checklist in the form of a short piece of text that supplied factual evidence in response to the question. Simple Yes or No questions were to be avoided. Where possible, information sources were to be cited. The MEFF information base is thus textual and qualitative.

## **b) The Scorecard**

6.6 The information in the checklists is scored with a simple traffic light system: green – the system or practice is in place; amber – it is under development; red – it is not in place. A blue score was entered if there was insufficient information available, and it was left blank if the question was not relevant to the agency. The scores were entered into a MEFF **scorecard**, which is an adaptation of the Balanced Scorecard format developed by Kaplan and Norton (1992). The scorecard provides three levels of information. At the most disaggregated level (level 3), it contains the scores for each question in the matrix. At level 2, the scores are aggregated within each organisational system and by perspective, by registering the average or predominant score in the category, with a small chip to reflect the presence of higher or lower scores. At level 1 they are aggregated by perspective. The Guidelines provided rules for attributing scores and aggregating them to levels 2 and 1. See figures 4 and 5, and annex 2.

6.7 The scorecard was developed with the assistance of an RBM specialist from the Centre for Management and Policy Studies (CMPS). He also organised two training events for MEG members, which introduced them to the Balanced Scorecard methodology and provided an opportunity to pilot the MEFF scorecard.

### **c) The Summary report**

6.8 A short, qualitative **summary report** provides an overview of the agency's strengths and weaknesses on each of the three perspectives, together with relevant background information on its structure and mandate, and any recent performance-related reforms. It also identifies three areas that will be used for future monitoring of the agency's effectiveness.

## **7. CONSULTATION PROCESS**

7.1 The DFID desk officers completed the first draft of the checklist and scorecard. There followed a period of consultation with the agencies on these drafts – more intensive in some cases than others – culminating in a final 'dialogue meeting', where discussions on the final draft were concluded. FAO was the only organisation that declined to participate formally in the MEFF exercise, although some staff assisted on an informal basis.

7.2 Each agency was provided with the full set of documents (approach paper, checklist and scorecard templates, guidelines). During the first three months, the lead IDAD adviser visited most of the agencies (20 in all) to explain what we were doing and why, and to provide an opportunity for discussion of any questions or concerns. The presentations were welcomed and allayed many of their concerns.

7.3 The multilaterals responded to our initiative in various ways: they awaited our drafts and then commented; they implemented the checklist themselves with a view to comparing versions (e.g. WFP, AfDB); or they completed it jointly with DFID staff (e.g. IFAD, UNIFEM). In many cases, we lacked sufficient information to complete the checklist internally and had to request information from the agencies. Sometimes this took the form of them sending relevant reports, other times they suggested text for the checklist.

7.4 The final 'dialogue meeting', was usually attended by senior management staff of the multilaterals (heads of department level) – an indication of how seriously they took the exercise. It involved detailed discussion of the textual answers as well as the scores. Although this was a consultation not a negotiation, in most cases it was easy to reach agreement. Undoubtedly, the agencies were concerned to maximise the scores, but many were very self-critical and actually reduced DFID's scores (see 9.22).

## **8. QUALITY ASSURANCE**

8.1 During the initial stages of implementation, quality assurance was provided informally through the MEG meetings, peer reviews within the ID

departments and one-to-one supervision by IDAD. More formal and systematic quality assurance (QA) commenced in late June with the appointment of an RBM specialist in IDAD. It focused mainly on consistency of application of the guidelines.

8.2 The QA process highlighted the need to standardise the format of the checklists, amplify the text in some cases of 'thin' answers, correct some errors in transferring scores from the checklists to the scorecards, and iron out some ambiguities in the rules for aggregation.<sup>10</sup> The changes resulting from this process involved reformatting the checklists, correcting scoring errors, expanding the text for some answers and modifying the scores at levels 1 and 2 according to the revised rules of aggregation (see Annex 2). The quality assured checklists and scorecards for 22 agencies were issued on 22 October 2004 pending the final review, and the UNESCO MEFF was added later.

8.3 A final step was to assess scoring consistency between assessments. We examined all the scores in the light of the evidence provided in the checklists and consistency of scoring across the 22 assessments. We found a small number of questionable scores (4.7% of all 1,574 scores), which were concentrated in only five agencies.<sup>11</sup> Questionable scores mostly arose on the boundary between green and amber scores, usually because of a tendency to consider not just whether a system was in place, but whether it was working well. (See section 9.20-9.24 below for further discussion)

8.4 In the interest of accuracy and cross-agency consistency, the questionable scores were adjusted. This mainly had the effect of raising scores, but there were also several cases where they were lowered.<sup>12</sup> Because of the small numbers involved, the overall impact of the adjustment is small both for individual agencies and in aggregate.

## 9. EVALUATION OF THE METHODOLOGY

### a) Evaluation of the instruments (checklist, scorecard, summary report)<sup>13</sup>

#### The checklist

*Was the checklist sufficiently generic?*

9.1 As already mentioned, we adapted the main checklist for humanitarian agencies to take account of the specificities of humanitarian work. However, even with this modification, unanticipated agency specificities created some problems in parts of the checklist. The main factors were:

---

<sup>10</sup> See section 9.25-9.28 below.

<sup>11</sup> Five agencies accounted for 65% of all such scores but these scores were only 13-17% of their total scores.

<sup>12</sup> 64% of score changes involved raising them, and 34% involved lowering them. The residual were movements between blue and blank scores.

<sup>13</sup> The analysis in this section refers to the initial 22 agencies that were completed by October. The UNESCO results came in during February 2005 and are not included here.

- Unusual governance arrangements limited the relevance of some corporate governance questions (EC, OCHA, OHCHR, UNAIDS, ICRC and IFRC);<sup>14</sup>
- Project-oriented operational questions (e.g. on project management, quality assurance and procurement) were not appropriate for coordination agencies (UNIFEM, IFRC, OCHA, UNAIDS);
- The global governance role of the Specialised Agencies was not covered;
- The presumption of development agencies working through government institutions was not relevant for EBRD, which lends primarily to the private sector;
- The presumption that effective organisations need to decentralise their operations to country level was not feasible for small agencies, such as UNIFEM, Habitat, UNIDO, IFAD;
- Inadequate emphasis was given to sub-regional programming compared with country programming, as the latter does not exist in some UN agencies (e.g. UNIFEM, FAO);
- Some of the questions on project financing (e.g. participation in SWAPs) were not appropriate for UN agencies;
- The low-income country orientation of the checklist was of limited relevance to some regional development banks (EBRD, AsDB, IADB), which lend primarily to middle-income countries;
- The generic relevance of support to PRSPs was questioned by some agencies.

9.2 The relevance of PRSPs was raised by a number of agencies: EBRD, which lends primarily to the private sector; advocacy and coordination organisations who see PRSPs to be about product and service delivery; and some Specialised Agencies, who consider that their global standard-setting role requires activities in developing countries that may not be included in PRSPs. DFID's position is that, as the entire international community has signed up to the principle of support to national PRSPs in those countries that have them, all agencies should consider how they relate to PRSPs even if this is a small part of their mandate and overall activity. In the absence of evidence of this consideration, the responses were scored red.

9.3 The issue of the generic applicability of the checklist was strongly challenged by some Specialised Agencies, not only on PRSP-related grounds, but regarding the perceived failure of the MEFF to cover their global governance role, much of which does not take place at country level (see Box 1). DFID will follow up these concerns with a more detailed study of how the effectiveness of the normative role of the Specialised Agencies should be assessed, and whether the MEFF should be adapted in the light of this. In the

---

<sup>14</sup> The EC has a complex intergovernmental process, plus two different directorates and a separate executing agency; OCHA and OHCHR are dependencies of the UN Secretariat and do not separate governing bodies; UNAIDS has a Program Coordinating Board consisting of UN co-sponsors and does not have direct intergovernmental representation; nor do the ICRC and IFRC which are private international NGOs.

meantime, we have assumed that since not all Specialised Agencies raised these concerns, the applicability of our questions was on the whole, valid.

**Box 1 Multilaterals' doubts about the generic applicability of the MEFF**

'The 26 agencies that DFID wanted to examine are not a homogenous group... it is a major oversimplification to attempt to measure their effectiveness using the same criteria. Specific questionnaires could have been developed for each homogenous group within the population without losing the capacity to compare the results. An opportunity has been lost by not doing so. The second dimension (country level results) was much too narrow in its focus for a specialised agency such as FAO. One must ask 'how can you measure the effectiveness of FAO if the structure of the questionnaire excludes the larger part of what it does (e.g. where is the Treaty on PGRFA or Codex or IPPC)?'

'One difficulty with such comprehensive assessments where the same questionnaire is being applied to some 26 organisations is that different terminologies are used in different organisations... Therefore, the questions have to be interpreted in a broader sense to see what is behind such a terminology. Furthermore, some questions may not be relevant due to the nature of the activities carried out by an organisation (e.g. difference between a funding organisation and a specialised agency such as UNIDO)'

'Thorough prior knowledge of OHCHR and the mechanisms of the UN Secretariat is required in order for the checklist to be objective. Often many of the questions were irrelevant or the complexity of our work was difficult to capture. The MEFF was more geared towards agencies with delegated authority.'

9.4 Questions that were considered to be not applicable on most of the grounds listed above were registered as blank scores. Only 3% of total scores were blank and they were concentrated particularly in two agencies – OCHA and OHCHR – which together account for two thirds of these scores.<sup>15</sup> These agencies have a number of special organisational features arising from the fact that they are part of the UN Secretariat and have limited autonomy over their governance, financing and evaluation arrangements.

9.5 Table 1 sets out these agency-specific factors in detail, including those relating to the humanitarian agencies. Question marks refer to the ambiguities over the relevance of PRSPs.<sup>16</sup> It provides a useful reminder of the importance of differences amongst multilaterals, not just regarding their mandate and function, but also their client base, types of partners, and size.

---

<sup>15</sup> However, blank scores were only a quarter of the total scores in these two agencies.

<sup>16</sup> They are included for the three regional development banks that mainly lend to middle income countries, although most of these now accept that they have to support PRSPs where relevant.

**Table 1. Limits to generic analysis across multilaterals**

Agency	Coordination agency, doesn't do projects	Global standards role	Size limits decentralisation	Small LDC involvement	Limited PRSP relevance	Unique governance structure	Limited involvement with government	Short term focus
<b>IFIs</b>								
World Bank								
AfDB								
AsDB				X	?			
IADB				X	?			
EBRD				X	?		X	
<b>UN F&amp;P</b>								
UNDP								
UNFPA								
UNICEF								
HABITAT			X					
UNIFEM			X					
<b>UN Sp Ag</b>								
UNIDO		X	X		?			
IFAD		?	X		?			
WHO		X			?			
FAO		X			?			
UNESCO		X			?			
<b>Humanitarian</b>								
UNHCR					X		X	X
ICRC					X	X	X	X
WFP					X		X	X?
IFRC	X				X	X	X	X
OCHA	X		X		X	X	X	X
<b>Other</b>								
UNAIDS	X	?	X		?	X		
OHCHR		X	X		?	X		
EC						X		

*What was the implication of having separate development and humanitarian checklists? Does it limit comparability between development and humanitarian agencies?*

9.6 As mentioned above the development checklist was modified because of the humanitarian agencies' more short-term focus, their particular funding arrangements (annual appeals), the fact that they do not work through governments or via government mechanisms such as PRSPs, and the specific nature of their inter-agency coordination arrangements. However, about 70% of the questions were common to both types of organisations, and a further 11% had very minor wording adjustments to refer more specifically to the humanitarian context. Thus about four fifths of the questions were the same or similar and provide a good basis for comparison between the two groups.

9.7 The remaining 19% of questions that were not applied to the humanitarian agencies were replaced by different questions for the latter group, but using similar criteria, e.g. results-focus, responsiveness to country context. Most of the adjustment occurred within the country-level results perspective, because of the removal of questions relating to MDGs, PRSPs, and delivering aid through government institutions. Instead, questions were asked about the availability of systems for assessing local needs, ensuring that delivery systems were appropriate for local conditions, appropriate staff training for humanitarian work, etc. Therefore, even though the precise questions differed in about a fifth of the questions, there is broad comparability in terms of the criteria being assessed.

9.8 The modification of the main checklist for the humanitarian agencies worked well, in that only a small proportion of answers were registered as 'not relevant' and most of these arose from factors unrelated to humanitarian work.<sup>17</sup> However, a minor complication arose from the fact that 7 of the questions that were deleted from the main checklist were not matched by new humanitarian questions, and five extra questions were added in other categories; so the number of questions within each category is not always the same. In all, the humanitarian checklist has 70 questions while the main one has 72. For the results analysis we have therefore compared average scores within categories rather than matching question for question.

9.9 It is important to note that the MEFF's coverage of humanitarian work is limited to five agencies for which this is their primary function. It does not cover the humanitarian work of development agencies such as UNICEF, UNDP, WHO and the EC. This raises the question of how the MEFF should deal with subsystems within agencies. This also arises in relation to the concessional wings of the development banks and the EC (IDA, AfDF, AsDF, EDF) and to Global Funds administered by multilaterals. At present, the MEFF applies to the whole organisation, not specific parts of it, on the grounds that the systems we are assessing are organisation-wide. However, there may be

---

<sup>17</sup> Most of the blank scores came from OCHA, and were related to its lack of involvement in direct project work because it is a coordination agency.

a case for looking more specifically at these subsystems in the contexts of ISs.

*Were the questions well formulated?*

9.10 We conducted a detailed analysis of the questions. This has revealed that a very small minority – 14 questions in all – were poorly formulated.

- Four questions produced all green scores across the 22 agencies – meaning that they failed to distinguish differences in agency performance.<sup>18</sup>
- Six questions asked whether a positive change was occurring, rather than whether there was a system in place. This may have produced more positive scores on these questions.<sup>19</sup>
- Four questions were insufficiently clear, and tended to produce answers that were not relevant to the issue in question.<sup>20</sup>

9.11 Broadly speaking, we were unhappy with the questions on corporate governance, strategy and human resource management. Although an organisation cannot be held accountable for the performance of its Governing Body, we were aware that weak or muddled governance can create difficulties for its performance.<sup>21</sup> However, there are few explicit standards for corporate governance of multilateral organisations, so it was difficult to establish assessment criteria for this component of the MEFF. The questions tended to focus on Board/Governing Body behaviour, as opposed to systems being in place, and it was difficult to get evidence for the views recorded. On the other hand, there were no questions on the effectiveness of senior management of the organisations, and some of the questions on governance would have been better formulated at this level (e.g. management of corporate risk).

9.12 The questions on corporate strategy were difficult to apply in cases where the agency did not have an identifiable strategy in place or had multiple strategy documents.<sup>22</sup> The partnership questions in this category – which were mainly formulated in terms of corporate commitments, were mostly non-contentious and produced green scores. The questions on human resource management were too general to identify adequately the factors underlying poor staff performance.

---

<sup>18</sup> These included two questions on partnership commitments in corporate strategies, a question on procurement of local inputs and a question on staff diversity policy.

<sup>19</sup> Most of these related to operational management (were the agencies simplifying their procedures, abandoning project management units, increasing programme selectivity, improving local level coordination?). The other questions were about increased strengthening of country programming and harmonisation of procurement.

<sup>20</sup> These included the questions on: corporate oversight of risk; the alignment of agency resources to PRS priorities; the peer review of country programmes (not specified that this was internal); and promotion to senior management (level not specified).

<sup>21</sup> For example, encouraging mission creep, giving mixed messages, etc.

<sup>22</sup> For example, IFAD, OHCHR

*How good was the response rate on individual questions?*

9.13 Of the 22 agencies included in this wave of assessments, almost all questions have been answered. 'No information' was recorded for only 19 questions out of a possible 1,574 – an infinitesimal proportion compared with most surveys.

*Were the answers sufficiently objective?*

9.14 We examined the textual responses to all 1,504 valid<sup>23</sup> answers in the 22 checklists. We could identify no case where the answer rested on a personal opinion or an impressionistic judgment. Subjectivity usually crept in where, in addition to the factual response, a comment is made about implementation or a suggestion is made as to what the agency should do, but this occurred in only 2% of answers and mainly in one agency.<sup>24</sup>

*Were the answers well evidenced?*

9.15 Our analysis indicates that a small proportion (7%) of the total valid answers could be said to have weak supporting evidence in the textual answer.<sup>25</sup> Weakly supported answers were concentrated in two agencies which accounted for 45% of this type of answers.<sup>26</sup> However, very few of the checklists had full referencing of sources. This was provided for only 7 agencies, with partial references for a further two. This was clearly a step too far for busy desk officers.

*How accurate are the answers?*

9.16 The accuracy of the answers is difficult to judge, given its textual nature and the lack of full referencing of sources. However, the multilaterals involved in discussions of the checklists were very keen to ensure that the factual information was accurate. The quality assurance process also picked up any inconsistencies between answers. Therefore we assume the accuracy of the data to be high.

*Overall, how useful was the checklist as an assessment tool?*

9.17 Our analysis suggests that the design of the checklists has been valid and its application robust. The small number of blank scores indicates that the criteria and questions have been sufficiently generic to apply to all but a small minority of advocacy and coordination organisations, and even there the majority of the questions were valid. The adapted humanitarian checklist permitted humanitarian agencies to be assessed on the same criteria as the

---

<sup>23</sup> Seventy 'no information' and 'not relevant' answers were excluded.

<sup>24</sup> In the checklists, parentheses have been inserted around such comments in the answers.

<sup>25</sup> Supporting evidence was considered to be adequate where several relevant paragraphs described the system in place. Weakly evidenced answers were excessively brief or did not directly answer the question. (Full referencing of information sources was excluded from this classification.)

<sup>26</sup> In these two agencies, weakly evidenced answers were about a third of total answers.

development organisations. Question completion was been very high and only a small number of questions were poorly formulated. However, the design could be improved with some fine-tuning of questions to improve applicability to coordination agencies and possibly to Specialised Agencies.

9.18 The vast majority of checklists answers have been objective and evidence-based, although without the full citation of sources that we would have liked. Some have been 'thin' on evidence, but they are a small minority. Consistency of application has been remarkable given the varied form of the textual inputs, the large number of desk officers involved (19 at any one time) and the variable degree of participation by the agencies. Much of this can be attributed to the ongoing discussions in the MEG, the guidelines and the quality assurance process.

9.19 Our overall conclusion, therefore, is that the checklists are robust and of good quality. The MEFF has achieved its basic objective of conducting assessments that are evidence based, factual and objective, although naturally, improvements can be made.

### **The Scorecard**

*How consistently were the scores applied?*

9.20 We have already discussed the integrity of the scores in the 8.3 above. To repeat, a detailed examination of the scores in relation to the evidence supplied revealed that only 4.7% of total scores were questionable. These scores were concentrated in a small number of agencies, and even there they were a minor part of their total scores.

9.21 Clearly the scoring of textual responses requires a judgment on the part of the assessor, and although the guidance for scoring was clear (the system is in place, it is not in place or a change is underway), there was still scope for ambiguity – particularly regarding whether or not a system is really in place, or whether an intended change was more than just a proposal. There may also be cases where a system is in place but is being further improved.

9.22 There was considerable scope for inconsistent scoring given the large number of desk officers completing the assessments. In view of the participative nature of the exercise, there could have been some implicit pressure to be generous with the scores. However, in many cases the harsher judgments were made by the agencies themselves; for example, UNHCR, WFP and UNESCO reduced some of DFID's green scores to amber.

9.23 As mentioned above, questionable scores mostly arose because of a tendency to consider not just whether a system was in place, but whether it was working well. Thus agencies were given an amber rather than a green score. Similarly, where a system was thought to be deficient it was given a red score, even though changes were underway. In one case, where the ambiguities lay on the boundary between amber and red, there was a tendency to give the agency the benefit of the doubt.

9.24 Given the inherent potential for subjectivity in the scoring process, the small proportion of questionable scores suggests that it was reasonably robust. However, even this small margin of inconsistency was considered unacceptable and the questionable scores were revised. We therefore consider that the final results are of a high standard of objectivity and consistency.

*Problems in aggregating scores to levels 1 and 2*

9.25 The rules for aggregating the MEFF ratings between levels 3, 2 and 1 were set out in the Guidelines. However, the quality assurance process revealed that these rules did not cover all possible situations and could result in undesirable biases. We therefore reviewed the situation, clarified any ambiguities and added rules to cover situations where there was none previously. Changes to the aggregated scores resulting from this decision were incorporated into the scorecards issued on October 22 2004.

9.26 The philosophy underlying the scorecard was that the scores should: a) provide a fair representation of the agency's performance, b) be an acceptable basis for dialogue, and c) provide an incentive for improvement. Therefore, the main emphasis was to be given to the predominant or average score, but the whole array was to be reflected by registering any score that was higher than or lower than the main score in small 'chips' in the upper left or right hand corners. Where it was not possible to identify a predominant or average score, i.e. where there was an equal split between two or four ratings, the lower score was registered because we wanted to emphasise where there was a need for improvement. Blank scores were to be excluded from the aggregation on the grounds that the aggregated scores should only reflect information that was relevant for effectiveness, but blue scores (i.e. no information on the question) were to be included on the grounds that it signalled that the information was potentially relevant for effectiveness but that we had none at this point.

9.27 The main problems with these rules related to: (a) the differential treatment of blue and blank scores, (b) the lack of an aggregation rule for when there was an even distribution between three scores; and (c) the left or right hand placement of the 'chips' could not follow the placement rule where the average score was a green or a red.<sup>27</sup> Under the revised aggregation rules, blue and blank scores would both be ignored in the aggregation; in cases of a 3-way split, the mid score would be reflected; the rule governing the placement of the 'chips' would not apply where the average or predominant score was red or green. In order to better reflect the full distribution of scores, it was also decided that level 1 would be aggregated directly from level 3. The revised rules are at Annex 2.

---

<sup>27</sup> A score above the main score was placed in the upper left hand corner, and one that was below it was placed in the upper right hand corner.

9.28 Checklist design further complicated the scoring system. First, the number of scores varied in different categories (two, three or four questions); second, where there was an even number of scores, it was difficult to determine the average or main score. Preferably, the checklist should have a constant and uneven number of scores in each category. Despite these difficulties, however, the traffic light system was vastly preferred to a numerical system both by the DFID desk officers and the agencies; both sides found it to be a more useful basis for dialogue.

### **The Summary Reports**

9.29 The quality assurance process focused primarily on the checklists and scorecards – a major effort in itself. The summary reports have received less attention so far. However, it is already evident that the style and quality of analysis in these reports is very variable. And the areas identified for monitoring lack indicators. Further work will be undertaken in these areas during the coming months.

#### **b) Evaluation of the overall approach**

*What does the MEFF tell us about multilateral effectiveness and what does it not? Is the RBM approach justified?*

9.30 The MEFF results indicate that the focus on organisational systems, using the three perspectives, has been able to discriminate the performance of the 22 agencies as well as yielding important and relevant information about their internal change processes (see Scott 2005). Feedback from the agencies has been very positive (see section 11 below).

9.31 However, it has to be recognised that asking whether a system or procedure is in place is not the same as asking about how well it works. The MEFF provides a starting point, but it needs to be complemented by information on quality and implementation. This would require in-depth studies or evaluations, which are time consuming, costly, and not amenable to annual monitoring. Rather than undertaking such work itself, DFID will urge the multilaterals to pay more attention to evaluating and reporting on the implementation of their performance systems. DFID will also collect information about multilateral performance at country level, through participation in the DAC, MOPAN and the SPA.<sup>28</sup> These multi-donor fora are engaging in a number of country level surveys on effectiveness issues.

## **10. TRANSACTIONS COSTS**

*Cost of developing the MEFF in-house.*

10.1 There have been costs and benefits to developing the MEFF in-house. DFID staff have had to undertake the work on a part-time basis, fitting it in

---

<sup>28</sup> These organisations have arranged surveys of donor/multilateral behaviour at country level. See OECD/DAC (2005b) and MOPAN (2005).

with their other tasks. For most of them, it was not anticipated in their annual workplans. They have had to bear the costs of innovation, with the inevitable problems that any learning-by-doing venture would entail (e.g. time spent in consultation, decision-making, revisions, etc). The work escalated in unexpected ways, and some continuity was lost because of staff turnover (see section 12.1). New desk officers who had not participated in the early developmental work or the initial training found it hard to understand what the MEFF was all about. Some people found the checklists and scorecards difficult to complete and we hugely overestimated our knowledge of the organisations. IDAD – a small and thinly stretched department – was simply unable to provide the necessary supervision and support to desk officers as they set about implementing the MEFF. Extra resources for quality assurance came late, when people were beginning to suffer MEFF fatigue.

10.2 The feedback questionnaires<sup>29</sup> indicate that some desk officers felt the MEFF to be time-consuming and difficult. Although they all agreed with the general approach, six of the ten thought it was too complex. Interestingly, this contrasts with the agencies' views.

**Box 2. DFID staff views about the MEFF checklists**

'It was not difficult (to complete the checklist) but required lots of concentration and hard work. Every box needs reflection, discussion and often two or three interlocutors feed into a single box. Information keeps coming in; I found I had to update around 6 or 7 times in all.

'It was very time consuming. In no case did we have enough information in-house. I found it hard to juggle working on these three checklists with my other areas of work... We tried different methods to get the information; in one case, we asked for the background documentation to be provided in order to form our own judgments. The amount of documentation provided was overwhelming and we simply did not have time to review it all'

'Completing the checklist was not too difficult but it was very time consuming to do it properly. One aspect was ensuring that we had the appropriate document reference to back up the comments in the checklist; this involved some micro detective work'

'This was extremely time-consuming as the initial completion of the task resulted in several blues (don't knows) and required extensive follow up with (the agency). This was somewhat protracted ...with both Geneva and New York HQ involved'

10.3 Unfortunately the time costs associated with the development and implementation of the MEFF were not logged at the time. However, it has been possible to reconstruct some of this information on the basis of

<sup>29</sup> Twelve questionnaires were returned in July/August 2004, but two were disregarded because they were from staff that had only a marginal involvement in the MEFF. Six desk officers failed to complete questionnaires. We have no way of knowing whether this is a biased sample.

retrospective calculations, records of meetings, travel and subsistence claims, training and other events, real-time monitoring of later MEFF activities (especially quality assurance), and estimates of IDAD overheads.<sup>30</sup>

10.4 This information is of variable quality, especially the retrospective calculations of the time spent by desk officers, which is likely to be underestimated. This element was largely based on a small number of feedback questionnaires, many of which lacked information on some activities, and did not cover activities after the questionnaires were sent in. Dialogue processes and subsequent checklist amendment that occurred afterwards were not covered. However, some of the IDAD work on late MEFFs was recorded in real time and entered into the calculations.<sup>31</sup> Other reasons for underestimation of desk officer time include the lack of estimates for some activities omitted from the feedback questionnaire (e.g. preparation of dialogue reports, feedback questionnaires) and loss of information because of staff turnover.

10.5 Table 2 covers the costs of developing and piloting the methodology from March 2003 up to the conclusion of the baseline at the end of October. It does not cover the results analysis, the review of the methodology and any other continuing work on the MEFF since October. Nor does it cover office inputs, such as paper, printing, telephone, meetings expenses etc.

10.6 The table shows that the overall cost of developing and implementing the MEFF was £193,811 or £8,810 per assessment. Total staff time was 812 person days or 37 days per assessment (just over seven working weeks). If we exclude the developmental work, we might expect a future complete agency assessment to cost £6,657 and 24 days of staff time, on average.<sup>32</sup> There would be cost savings as quality assurance and training were routinised, but there would be extra costs relating to dissemination and further work on links to Institutional Strategies and PSA monitoring.

10.7 IDAD has absorbed 70% of the costs of MEFF, mostly through staff costs on design, coordination, quality assurance and external presentations.<sup>33</sup> IDAD's travel costs associated with agency presentations and MEFF dialogue with the development banks were also substantial.

---

<sup>30</sup> The basic unit of calculation was hours, which was converted to staff costs based on the standard DFID unit cost per grade.

<sup>31</sup> The number of hours was averaged according to the number of responses and then scaled up for the 22 assessments.

<sup>32</sup> This figure includes IDAD overhead costs such as coordination, quality assurance and training. A complete update of the baseline data is not anticipated for 3 years.

<sup>33</sup> 50% of a Senior Adviser over 20 months, 80% of a policy adviser over five and a half months, 50% of a research assistant over 12 months, and clerical work

**Table 2 TRANSACTIONS COSTS OF THE MEFF**

	days	£	%
<b>Methodology development</b>			
MEG meetings	27.0	4,953.68	
Agency consultations	17.0	4,519.28	
Piloting checklists	17.4	3,065.33	
CMPS consultancy 10 days @£500	10.0	5,000.00	
Research assistance	130.0	7,200.00	
SSDA residual time (10 months@50%)*	86.1	22,621.72	
<b>TOTAL</b>	<b>287.5</b>	<b>47,360.01</b>	<b>24.4</b>
<b>Assessments</b>			
Initial assessments	50.1	8,808.80	
Agency presentations	24.9	29,407.79	
Internal peer review	17.1	3,779.94	
Peer-review follow up	11.8	2,081.20	
Desk communications with agencies	11.9	2,097.33	
Dialogue with agency	19.3	3,388.00	
Dialogue follow up	47.4	8,335.56	
Summary report	18.2	3,199.78	
Travel and subsistence	41.3	18,956.00	
SSDA residual time (10 months@50%)	86.1	22,621.72	
MEG meetings	27.0	4,953.68	
Other	31.6	5,566.00	
<b>TOTAL</b>	<b>386.6</b>	<b>113,195.80</b>	<b>58.4</b>
<b>Quality assurance</b>			
Analysis of checklists	10.3	2,101.66	
QA meetings	9.8	1,599.74	
QA follow up (clerical)	33.8	5,366.50	
PA residual time (5.5 months@80%)	30.8	12,824.47	
<b>TOTAL</b>	<b>84.6</b>	<b>21,892.37</b>	<b>11.3</b>
<b>Training</b>			
CPMS	5.0	2,500.00	
Staff time	40.0	5,643.23	
Staff travel & subs.	5.3	2,464.00	
SSDA	0.8	199.37	
<b>TOTAL</b>	<b>51.0</b>	<b>10,806.60</b>	<b>5.6</b>
<b>Other</b>			
General communications	2.0	531.66	
Other		24.99	
<b>TOTAL</b>	<b>2.0</b>	<b>556.65</b>	<b>0.3</b>
<b>GRAND TOTAL</b>	<b>811.7</b>	<b>193,811.43</b>	<b>100.0</b>

---

\* excludes time calculated into other activities

10.8 If we convert the estimate of staff time into a consultancy daily rate of £500, the estimated cost of contracting this work out would have been more than double, i.e. £406,000 in total and an average cost per MEFF assessment of £18,454.<sup>34</sup> And there would doubtless have been extra costs for DFID staff in terms of inputs to the consultancy. The financial costs of undertaking the MEFF assessments in house have therefore been low, relative to an external operation, and have mostly been absorbed because of synergies with other ID work. However, there has been a significant cost in terms of delays to completing the work, with knock on implications for new Institutional Strategies and PSA monitoring.

10.9 Total MEFF in-house costs could probably have been lower if the quality assurance process had been more formally set up at an earlier stage and if IDAD had been better resourced to support the desk officers. This would have avoided the extra work generated by late changes. They would also have been lower if there had not been such high staff turnover in the UN and Commonwealth Department.

#### *Benefits of conducting the MEFF in-house*

10.10 These costs are more than offset by the very considerable benefits of undertaking the MEFF assessments in-house. First, there has been an immense internal learning process about the organisations and widespread ownership of the results. All desk officers said that their understanding of their agencies – as well as of RBM and of effectiveness issues generally – had significantly improved. This will contribute to better management of ISs and PSA monitoring in the future, in which the desk officers play the key role. These benefits would not have been achieved if the MEFF had been contracted out to consultants. There would have less understanding of the issues and less ownership of the results.

#### **Box 3. MEFF as a learning tool**

'We now have a more solid understanding and evidence of their work and efforts'

'I found out things I never knew about my agencies'

'The process showed a number of gaps in my knowledge of the organisations'

'Very good as introductory briefing material on the agencies'

10.11 A second major benefit is the credibility of the MEFF with the agencies that came from the personal involvement of DFID staff. The participation of the agencies implied transactions costs for them and it might be questioned whether this would have been forthcoming if the information had been requested by consultants.

---

<sup>34</sup> This figure would exclude any overheads and incidental expenses that a company might charge.

## *Costs and benefits of the participative approach*

10.12 Despite our good intentions, the MEFF was not cost-free for the agencies: we relied on them enormously to supply much of the checklist information.<sup>35</sup> The agency consultations were therefore not just about transparency and good public relations, they played a crucial role in enabling the checklist to be completed, evidence-based and factual. The joint scrutiny of the textual input to the checklist helped to strengthen the factual emphasis of the answers, providing a challenge to any subjective judgement that could not be justified. The agency consultations also played a major role in securing ownership of the results by the agencies and their support for the initiative as a whole (see reports from desk officers in box 4, and box 5).

### **Box 4. Consultative approach improves accuracy and evidence-base**

‘WHO expressed its commitment to MEFF and said they appreciated the opportunity to discuss and influence the results as they had concerns about some of the substance. Several inaccuracies had been noted in the draft provided, so drafting changes and corrections would be offered.’

‘A number of scores had to be changed from the first draft and we were able to add a significant amount to the explanation of the ratings. A number of the original draft ratings were clearly wrong, with insufficient information being available in London to make sound judgments.’

‘We went through the checklist column by column checking the accuracy of the responses and identifying additional information where needed. Several changes were required. One of the managers had brought written text in response to the checklist questions and several documents were sent to us afterwards.’

10.13 However, the consultations did result in delays to completing the process, as requests for information and meetings had to fit in with their timetables as well as ours. There is also the possibility that the dialogue process may have influenced some of the scores, but we are confident that the quality assurance process has mostly corrected this tendency.

## **11. VIEWS OF THE MULTILATERALS**

11.1 Feedback from the agencies<sup>36</sup> indicates extensive support for the MEFF approach, both in terms of the systems focus, the three perspectives and (on the whole) the indicators used. The approach and methodology was considered to be appropriate; two agencies (IFAD and AfDB) suggested that the MEFF should become the industry standard for multilateral effectiveness assessments. Some agencies said that they intended to use the methodology and some of the indicators for their own internal processes.

<sup>35</sup> Early internal drafts of some checklists had a large number of ‘no information’ responses.

<sup>36</sup> Feedback questionnaires were received from 16 of the assessed agencies. Nil responses included EBRD, OCHA, ICRC, IFRC, AsDB and Habitat.

11.2 The agency feedback questionnaires indicate that 14 of the respondents thought that we had adopted the right approach, thirteen thought that the results were sufficiently evidenced based, and more than half of them thought that the checklist was not too complex, not too superficial and was sufficiently objective. A few (e.g. FAO, OHCHR and UNIFEM) had concerns about whether we had the right indicators. Almost all were satisfied with how we had handled the process (transparency and consultation). Most thought we had addressed their concerns. Eleven said that the MEFF had helped their internal discussions on effectiveness and fourteen said it was valuable to have an external view. Nine said that they didn't mind the transactions costs involved. These results are set out at Annex 3.

11.3 Some comments on the general value of the exercise are at box 5.

**Box 5. Multilateral views on the usefulness of the MEFF**

'This was a productive exercise, conducted in a professional and cooperative way, in a positive, creative and consultative environment' (World Bank)

'We welcome this assessment very much as in the process it has helped us to be more self-critical, as well as given us the confidence to continue to build on the institutional reforms that will make the ADB truly more effective as a development partner in the delivery of aid to our RMCs.' (AfDB)

'We very much welcome the approach that has been used, which builds directly on our own organisational systems for measuring performance and assessing impact.

'As UNIFEM is moving forward in refining its own organisational effectiveness matrix (the MEFF) is extremely helpful'.

'We welcome such occasional independent assessments as they, inter alia, encourage an organisation to look at its operations in a holistic manner, provide the organisation with added assurance on its mode of operations, highlight its strengths and relevance, as well as those areas requiring more attention. Such assessments also provide additional assurance to various stakeholders. (UNIDO)

'On the whole, we have found the exercise to be useful, especially as it gives us an external donor perspective on the programme.' (UNAIDS)

'Many participants to the MEFF process on the IDB side found it to be a valuable exercise' (IADB)

11.4 Comments on the general methodology were positive, with the exception of the reservations about generic applicability already referred to above (paragraph 9.3 and box 1).

**Box 6. On the validity of the MEFF approach and design**

‘The three component approach used in this framework, namely focusing on country results, internal performance and partnerships, is broadly consistent with the approach adopted by the new MYFF. The checklist provides a baseline on processes, systems and performance measurement instruments.’ (UNDP)

‘The focus on these aspects (perspectives), in addition to being substantive as an approach also provides a balanced perspective both on the current performance and potential impact of an institution such as ours. (AfDB)

‘The overall approach was very helpful to us, particularly the focus and level of questions. For UNIFEM, there was an over-emphasis on country-level results, given that we operate at the sub-regional level. This is a particularly important level for promoting learning and south-south exchange and could be better captured and valued with the addition of some questions. Based on first-time use, it was a very helpful ‘prod’ to stimulate thinking about different elements of organisational effectiveness.’

‘Broadly speaking, the application of the balanced scorecard was an interesting and stimulating way of approaching a complex problem.’ (FAO)

11.5 The agencies’ appreciation of the consultative process is illustrated below:

**Box 7. Views on how we handled the process**

‘We greatly appreciate the participation and open process that you have taken for the preparation of the balanced scorecard. We at UNFPA are learning from this interesting experience’. (UNFPA)

‘The process showed a high degree of collaboration, trust and mutuality. UNIFEM would want to convey great appreciation for the process, which is consistent with DFID’s overall approach of valuing learning and self-assessment as a part of enhancing effectiveness’ (UNIFEM)

‘We very much appreciated the partnership approach followed by DFID for this exercise and the professionalism displayed by those involved in the process’ (UNIDO)

‘We were satisfied with the fairness, degree of consensus and consultation, including on areas of weakness.’ (AfDB)

11.6 The major concerns raised by the agencies related to the need for bilaterals to coordinate amongst themselves on exercises such as this.

**Box 8. On the need for bilaterals to coordinate such initiatives**

‘There needs to be coordination amongst bilaterals interested in assessing MDB performance’ (IDB)

‘We think it should be a concerted and unified exercise of all the bilaterals together in the interest of harmonisation and to avoid duplication of effort and assessment fatigue on the part of development agencies’. (AfDB)

‘We could not afford to do this exercise for all donors’ (UNESCO)

‘More bilateral organisations should ideally join this effort simultaneously’ (IFAD)

‘An agreement with all EB members and donor partners on which indicators should be used ...could make the process of regular performance reviews more efficient’ (WFP)

‘It is important that such assessments/evaluations are coordinated with other stakeholders (e.g. with other Member States), and that the results are shared with all concerned for the sake of transparency. This, inter alia, had the advantage of including other stakeholders as partners in the process, thus reducing the necessity for and costs of similar reviews by other Member States.’ (UNIDO)

‘There is a concern about having multiple frameworks to report our organisational effectiveness beyond the agreed normal reporting through the MYFF reports and annual report to the Executive Board. Although we do not know DFID’s plans for reporting on the MEFF, we would like to stress the higher transactions costs that may be created if additional reporting will be required’ (UNFPA)

## 12. RISKS AND SUSTAINABILITY

12.1 The main internal risk to the sustainability of the MEFF is related to staff turnover, which undermines the main benefit of conducting it in house: the knowledge and learning referred to in 10.10. The MEFF required inputs from a core of 19 desk officers who lead on the relevant ISs, plus 2 advisers in IDAD. Of these 21, only 8 are now in post and 4 are due to move on in early 2005. This will represent an 81% loss of MEFF experience over a period of two years. This high figure is partly due to the relocation of the UN and Commonwealth Department to Scotland in mid 2004, which resulted in 100% staff turnover. A total of 35 staff have been involved in the MEFF at some time and the majority of the current IS desk officers have not been directly involved in completing the MEFF instruments. This internal turnover will require substantial investment in staff training and follow up advice from IDAD. There are issues about whether ID has the resources to do this from its current advisory pool.

12.2 Another internal risk is that DFID staff in other Divisions, including senior management, who have little awareness of details of the MEFF, express different views about the effectiveness of some multilaterals, thus undermining the MEFF's credibility with the agencies. Therefore substantial internal dissemination of the MEFF is required to ensure more widespread ownership and understanding.

12.3 The external risk is that of non-cooperation on the part of the agencies. Because DFID has played the part of innovator in this field, we have benefited from their collaboration. But already they are expressing reluctance to participate in more ventures of this sort, especially as other bilaterals embark on similar initiatives. So it may be more difficult to update the MEFF in three years time. In order to address this risk, DFID needs to share the MEFF widely with other bilaterals and advocate a joined up approach to multilateral assessment. We are already doing this to a degree through MOPAN and the DAC, although we have to be aware that joint initiatives also involve substantial transactions costs.

### **Next steps**

12.4 The main task for the next two years will be to focus on monitoring the three focus areas as part of the IS and PSA reporting. Follow-up work is required to establish guidance on how the MEFF should be integrated with the ISs and to identify indicators for the three MEFF monitoring areas. A full revision of the baseline will take place in 3-4 years time, unless an agency requests one earlier.

12.5 In addition, further investigative work is recommended to consider how to assess (a) the global standards role of the Specialised Agencies, and (b) subsystems within multilateral agencies (e.g. concessional arms of MDB, Global Funds, humanitarian departments of development agencies, development work of humanitarian agencies, etc).

12.6 Full institutionalisation of the MEFF will require:

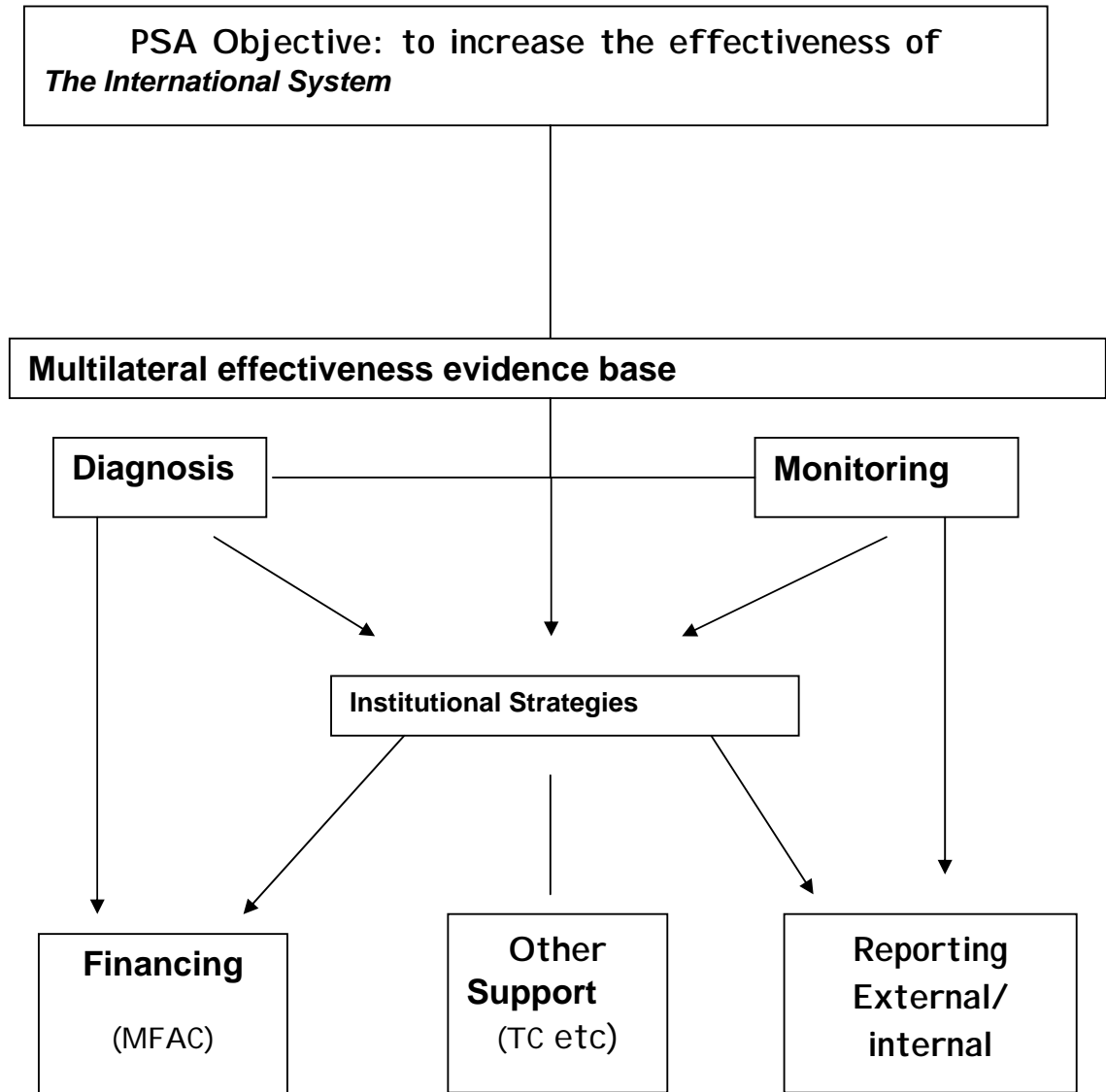
- An internal training programme for the desk officers and their managers
- The integration of MEFF work into the workplans and performance appraisals of IS desk officers
- Adequate resources for coordination, supervision and support in IDAD
- An internal dissemination programme to other Divisions, country programmes and senior management
- An external dissemination programme for Whitehall and bilateral agencies.

**END**

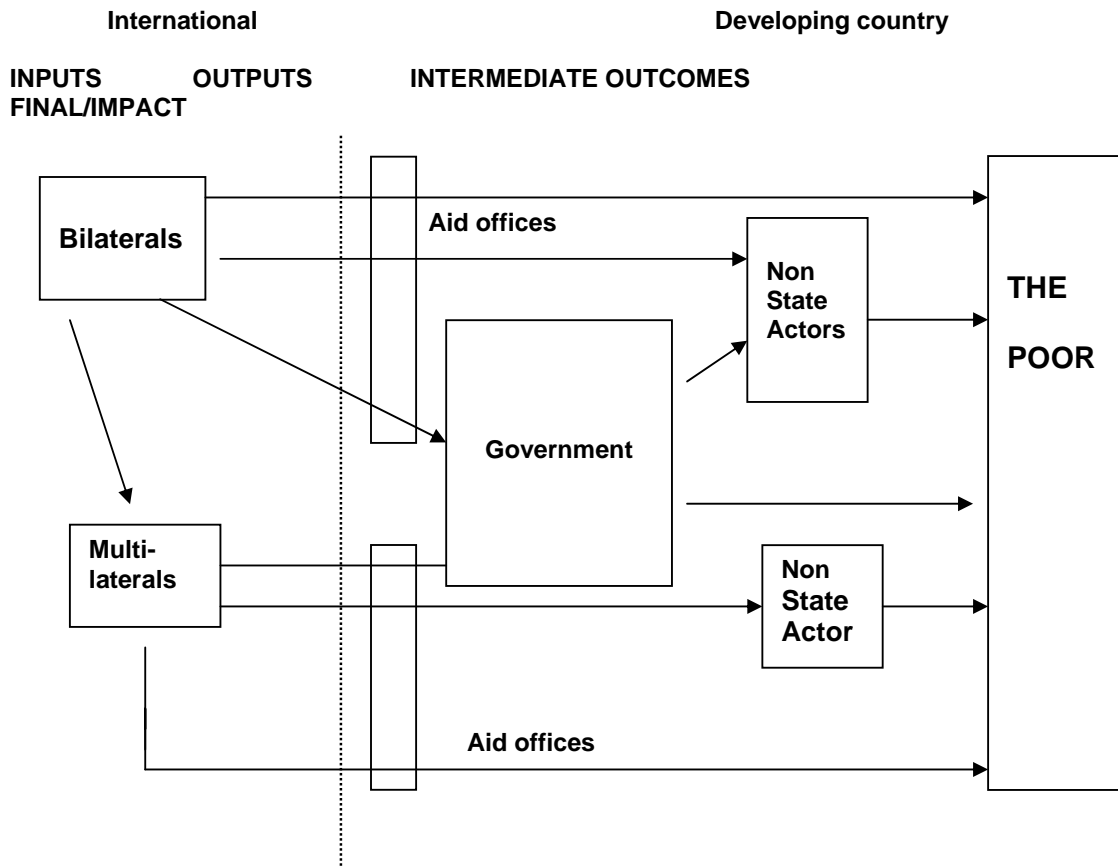
# **THE MEFF METHODOLOGY**

## **ANNEXES**

Figure 1. Multilateral Effectiveness Work



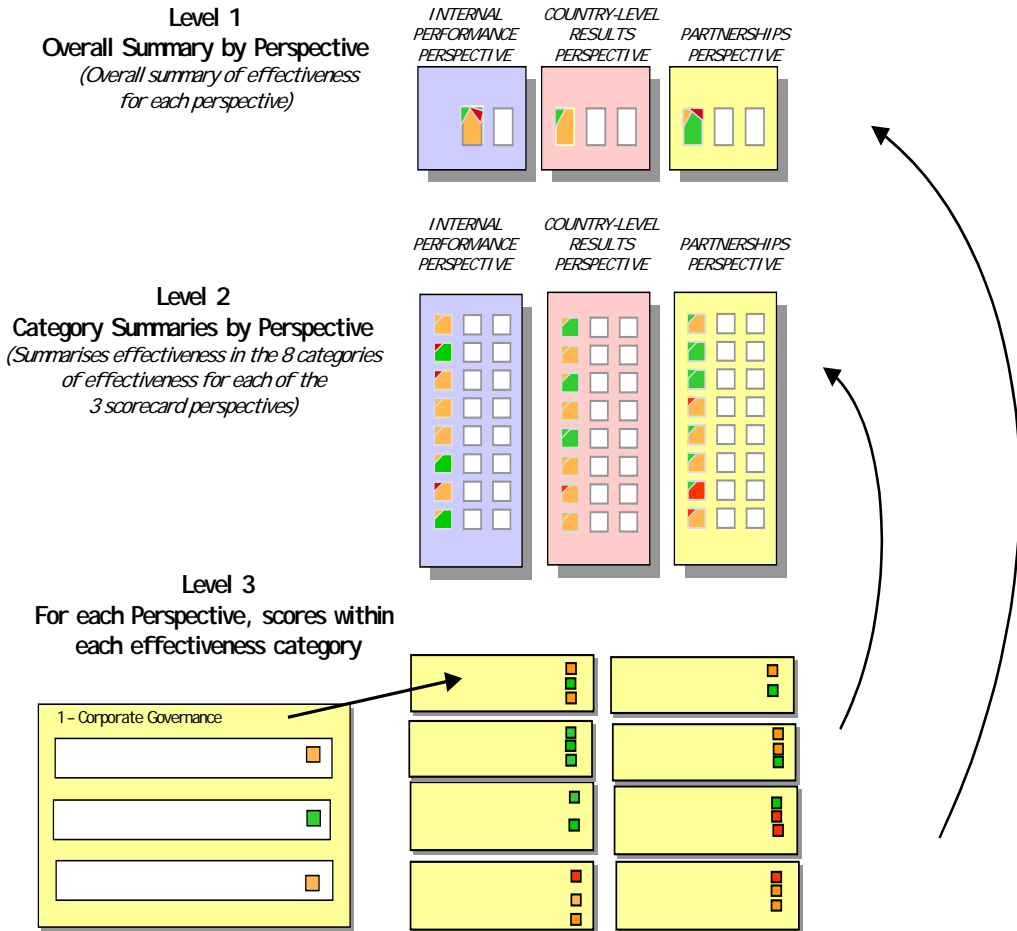
**Figure 2. Aid relationships and the results chain**



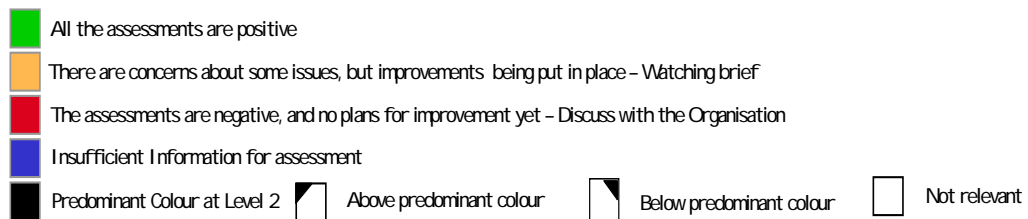
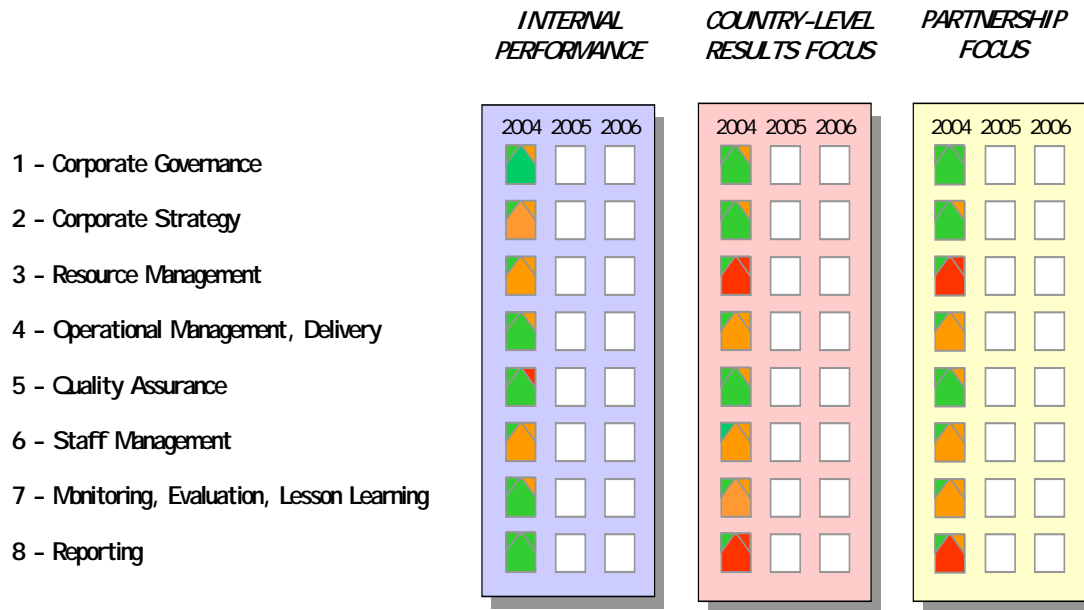
**Figure 3. The MEFF checklist**

Organizational systems	Internal Performance	Focus on country Level results	Focus on partnership
Corporate Governance	3 questions	3 questions	3 questions
Corporate Strategy	3 questions	3 questions	3 questions
Resource Management	4 questions	3 questions	2 questions
Operational Management	4 questions	2 questions	3 questions
Quality Assurance	4 questions	3 questions	2 questions
Staff Management	4 questions	2 questions	3 questions
M&E Lesson learning	3 questions	3 questions	3 questions
Reporting	3 questions	3 questions	3 questions

Figure 4. MULTI LATERAL EFFECTIVENESS SCORECARD



**Fig. 5 MULTILATERAL EFFECTIVENESS SCORECARD**  
**Dream Development Agency - Level 2 Summary**



## Aggregating the MEFF scores

The rules for aggregating the MEFF scores between levels 3, 2 and 1 were set out in the Guidelines. However, the quality assurance process revealed that these rules did not cover all possible situations and could result in undesirable biases. We have therefore reviewed the situation and clarified any ambiguities and added rules to cover situations where there was none previously. The new rules are set out below.

### 1. General principles used in constructing these aggregation rules:

The philosophy underlying the scorecard is that the scores/scores should: a) provide a fair representation of the agency's performance, b) be an acceptable basis for dialogue, and c) provide an incentive for improvement. Therefore:

- The main emphasis is given to the predominant or average score
- Where it is not possible to identify a predominant or average score, i.e. where there is an equal split between two or four scores, the lower score is registered because we want to emphasise where there was a need for improvement
- The whole array is reflected by registering any score that is higher than or lower than the predominant or average score by adding 'chips' at the top of the box
- Blue (i.e. question was not relevant for the organisation) and blank (i.e. no information on the question) scores are recorded at level 3 because they provide important information about question coverage.
- Blue and Blank scores are excluded from the aggregation on the grounds that the summary scores should only reflect information that was relevant for effectiveness

### 2. Revised aggregation rules

- Identify the predominant or average (mid point) colour for each variable group on a simple arithmetic basis and fill this in the main area of the score box,
- Where there is an equal two-way split, register the lower colour
- Where there is a score *above* the predominant colour, register this in the small 'chip' on the *left hand side* of the box,
- Where there is a score *below* the predominant colour, register this in the small 'chip' on the *right hand side* of the score box,
- Where the predominant colour is at the extreme of the spectrum, i.e. green or red, and the full set of scores is registered, arrange the chips in the same left-right order, even though they may be higher or lower than the predominant colour.
- Ignore blue and blank scores in the aggregation

### 3. Aggregating to level 1

Aggregate to level 1 from level 3, on the basis of the full set of scores within each perspective, according to the same rules as above.

IDAD  
12/10/04

**DFID MEFF AGENCY FEEDBACK QUESTIONNAIRE**

Thank you for participating in the dialogue on DFID’s assessment of multilateral effectiveness. We would greatly value your feedback on the process, which should only take a few moments. The questions below have a scale illustrated by opposing statements. Please tick one point in each scale. The questionnaire, and any more extensive comments you may wish to make, can be handed back to DFID staff or emailed to [a-scott@dfid.gov.uk](mailto:a-scott@dfid.gov.uk)

Name of your agency..... Date.....

1. Please comment on the substantive approach we have adopted: the focus on organisational systems; their internal performance, focus on country results, partnerships:

Broadly the right approach    **8**    **6**    **2**    —    —                      Not the right approach

2. Do you have any reservations about the checklist? On the whole, do you think it is

Too complex	—	<b>4</b>	<b>4</b>	<b>4</b>	<b>4</b>	Not too complex
Too superficial	—	<b>2</b>	<b>3</b>	<b>3</b>	<b>8</b>	Not too superficial
The right indicators	<b>3</b>	<b>4</b>	<b>5</b>	<b>3</b>	<b>1</b>	Not the right indicators
Too subjective	—	<b>1</b>	<b>4</b>	<b>6</b>	<b>5</b>	Sufficiently objective

3. Do you think that the results are sufficiently evidenced-based?

Yes    **13**    **1**    **2**    No

4. How satisfied are you with the way we have handled the process?

Transparency:	Satisfied	<b>14</b>	<b>1</b>	<b>1</b>	—	—	Not satisfied
Consultation:	Satisfied	<b>12</b>	<b>1</b>	<b>3</b>	—	—	Not satisfied
Addressed our concerns:	Satisfied	<b>10</b>	<b>3</b>	<b>2</b>	—	<b>1</b>	Not satisfied

5. Has the process benefited you in any way?

It helped our internal discussions	<b>8</b>	<b>3</b>	<b>2</b>	—	<b>3</b>	did not help
It was valuable to have an external view	<b>11</b>	<b>3</b>	<b>1</b>	<b>1</b>	—	not valuable
It implied transaction costs we can ill afford	<b>2</b>	<b>2</b>	<b>3</b>	<b>3</b>	<b>6</b>	we don’t mind

6. Do you think that bilaterals should be involved in this kind of exercise?

Yes    **7**    **5**    **2**    No

7. Do you have any further comments, concerns or recommendations? (please comment below)

Thank you for your cooperation

## REFERENCES

- Balogun, P. (2003) *Reporting Multilateral Effectiveness*, report for IDAD, DFID. March
- Bezanson, K, et. al. (2003) *Evaluating Development effectiveness in Six United Nations Agencies*, IDS, Sussex, April
- Binnendijk, A. (1999) *Results Based Management in the Development Co-operation Agencies: a Review of Experience*, Background Document No 3, OECD-DAC Working Party on Aid Evaluation, November
- Dyer, N. et. al. (2003) *Strategic review of Resource Allocation Priorities*, DFID Discussion Paper, January.
- Farquhar, C.R. (2000) *Governments Get Focused on Results: Integscore Performance Measurement into Management Decision Making*, The Conference Board of Canada
- Flint, M. (2003) *Easier Said Than Done: A Review of Results-based Management in Multilateral development institutions*, Report for IDAD, DFID. March
- Kaplan, R.S. and Norton, D. P. (1992) *'The Balanced Scorecard – Measures that Drive Performance'*, Harvard Business Review, Jan-Feb 1992: pp 71-79.
- Lusthaus, C. et. al. (2002) *Organisational Assessment: A framework for Improving performance*, IADB/IDRC, Washington D.C. and Ottawa.
- Meier, W. (2003) *Results-Based Management: Towards a Common Understanding among Development Cooperation Agencies*. Discussion Paper prepared for CIDA and DAC Working Party on Aid Effectiveness.
- MOPAN (Multilateral Organisations Performance Assessment Network) (2005) *The MOPAN Survey 2004. Synthesis Report*. January
- National Audit Office (2002) *Performance Management – Helping to Reduce World Poverty*, April
- OECD/DAC (2005a) *Results-Based Country Programming: Improving Aid Agencies Performance in Managing for Development Results*. Draft Emerging practices Note. Task Team on Agency Performance of the Joint Venture on Managing for Development Results.
- OECD/DAC (2005b) *Survey on Harmonisation and Alignment: Progress in Implementing Harmonisation and Alignment in 14 partner Countries*.
- Schacter, M. (1999) *Results-Based Management and Multilateral Programming at CIDA: A Discussion Paper*, Institute on Governance, Ottawa. November
- Scott, A. (2005) *DFID's Assessment of Multilateral Organisational Effectiveness: An Overview of Results*. DFID, March 16.

Scott, A. (2004) *Assessing Multilateral Effectiveness*. IDAD, DFID, February

Turner, R. et al., (2003) *Vision and Options for Change for the International Development Architecture*. IDAD, DFID. July.

World Bank (2002) *Better Measuring, Monitoring and Managing for Development Results*, Paper for Development Committee, September