reports

DFID Department for International Development

EVALUATION REPORT EV645

November 2003

Measuring Value for Money ?

An independent review of DFID's Value for Money (VFM) Indicator, Public Service Agreement 2003-2006

By Derek Poate and Christopher Barnett

DEPARTMENT FOR INTERNATIONAL DEVELOPMENT

EVALUATION REPORT EV645

MEASURING VALUE FOR MONEY?

An Independent Review of DFID's Value for Money (VFM) Indicator, Public Service Agreement 2003–2006

Derek Poate and
Christopher Barnett

The opinions expressed in this report are those of the authors and do not necessarily represent the views of the Department for International Development.

PREFACE

This study was undertaken as part of the programme of independent evaluation studies commissioned by the Evaluation Department of the Department for International Development (DFID). The purpose of these studies is to improve the quality of development activities by providing evidence of what makes for effective development. Lessons learned from evaluation can be applied to strengthen current and future policies and programmes. Evaluation of development assistance provided by DFID also helps to strengthen DFID's accountability.

DFID's Evaluation Department (EvD) is independent of the spending divisions in DFID, and reports to DFID's Management Board through the Director General (Corporate Performance and Knowledge Sharing). Each year, Evaluation Department commissions a number of evaluations which rigorously examine the design, implementation and results of selected DFID policies and programmes. The findings and lessons from each evaluation are published.

This report presents findings and conclusions from an independent assessment of the Value for Money indicator, which constitutes an important element of DFID's Public Service Agreement. The Public Service Agreement sets out DFID's corporate objectives and targets, which define the contribution DFID aims to make towards achieving the global Millennium Development Goals. Unlike most of the PSA objectives, which focus on progress towards the Millennium Development Goals, the Value for Money indicator is directly linked to the effectiveness of activities funded by DFID. It is therefore an important measure of DFID's operational performance.

The study was conducted by Derek Poate and Christopher Barnett of ITAD Ltd. This was undertaken largely a desk study, supplemented by interviews and internal consultations. Further work is planned to test some of the conclusions.

DFID's management response is included at Annex 6. Annex 7 presents data on scores for achievement and risks at the time of going to press.

The study was managed by Joanne Asquith and edited by Caryn McLean.

Colin Kirk
Head, Evaluation Department
November 2003

Preface

CONTENTS

ABBREVIATIONS AND ACRONYMS

APPR	Annual Plan and Performance Review
ARDE	Annual Review of Development Effectiveness
ARPP	Annual Review of Portfolio Performance
CAP	Country Assistance Plans
CHAD	Conflict and Humanitarian Affairs Department
DAC	Development Assistance Committee
DBS	Direct Budgetary Support
DER	Development Effectiveness Report
DFID	Department for International Development
EAPD	Eastern Asia and Pacific Department
EEWH	Eastern Europe and Western Hemisphere Department
EMAD	Europe, Middle East and Americas Division
FAO	Food and Agriculture Organisation of the United Nations
ICR	Implementation Completion Report
MIS	Management Information System
MTR	Mid-Term Review
NAO	National Audit Office
OECD	Organisation for Economic Co-operation and Development
OI	Office Instructions
OPR	Output to Purpose Reviews
OVI	Objectively Verifiable Indicators
PAD	Project Appraisal Document
PCR	Project Completion Reports
PPCM	Programme & Project Cycle Management
PRISM	Performance Reporting Information System for Management
PSA	Public Service Agreements
PSR	Project Status Report
QAE	Quality at Entry Assessment
QAG	Quality Assurance Group
QER	Quality Enhancement Review
QQT	Quantity, Quality and Timing
QSR	Quality of Supervision Review
SDA	Service Delivery Agreement
SWAps	Sector Wide Approaches
UNESCO	United Nations Educational, Scientific and Cultural Organisation
VFM	Value for Money

1. KEY FINDINGS AND CONCLUSIONS

The VFM indicator captures a large proportion of DFID's bilateral expenditure but there are important concerns, some of which may need to be addressed for subsequent Public Service Agreements (PSAs). Key concerns include:

- The criteria for eligibility excludes 84% of projects[1] and it is not clear how important these operations are for DFID's poverty objectives, nor the extent to which the same types of projects are excluded in each region.

- There are apparent regional anomalies in scoring, especially across risk categories.

- The current system is not particularly well suited to measuring the performance of Sector Wide Approaches (SWAps) and Direct Budgetary Support (DBS).

Supporting evidence and quality of Output to Purpose Reviews (OPR) is reasonably thorough, with innovative approaches being developed in Country Offices, but:

- Only a small proportion of review documents are available from the Performance Reporting Information System for Management (PRISM)[2].

- There are many gaps in the supporting narrative[3].

- The style and content of the text varies widely.

The scores in the system show a tendency to clustering and clear anomalies in the treatment of risk. This may arise from the nature of the scoring system or the limited guidance given to staff. At present, processes that are thorough in each country may yield results that are contradictory in aggregate.

There is evidence that the scoring system has settled down over a relatively short time. Despite the small samples studied in this review, the issues that have emerged are important if there is to be confidence in the robustness of the indicator.

Summaries of the main findings can be found in boxed text at the end of each section, and a detailed list of the key issues is provided in section 8 (Conclusions and Key Issues).

[1] This is according to an analysis of PRISM data for operational projects, undertaken in January 2003. However, only 12% of operational projects are excluded by commitment value.
[2] Only a small proportion of the original review documents (OPRs, Project Completion Reports (PCRs), etc.) are uploaded onto PRISM. This is distinct from the summary information (scores, justifications, etc.) which is entered directly onto the system.
[3] It is not presently mandatory to complete all the text fields.

2. INTRODUCTION

This study examines the use of scoring by DFID to report performance against the PSA Value for Money (VFM) indicator. The indicator[4] has two parts: an objective for increasing overall value for money and a measure of progress against it.

Value for money under the 2003–2006 PSA is achieved in DFID when:

1. The proportion of DFID's bilateral programme going to low income countries increases from 78% to 90%; and

2. There is a sustained increase in the index of DFID's bilateral projects' evaluated success.

The first part is measured from records of new commitments. The second part is based on scored assessments of performance by DFID staff. This latter part is the subject of the review.

Three recent studies have drawn attention to limitations in current arrangements for project and programme monitoring. A peer review by the Organisation for Economic Co-operation and Development/Development Assistance Committee (OECD/DAC) in 2001 found that monitoring and evaluation of portfolio performance had little ownership by DFID staff, resulting in low compliance rates.[5] The study identified a need to reconcile the targets in the PSA with DFID's longer-term development objectives. The DFID Development Effectiveness Report raised concerns about the use of self-assessment, concluding that operational staff need more support to achieve adequate coverage, consistency, timeliness and quality of reporting.[6] A National Audit Office (NAO) review in 2002 found strengths in DFID's approach to performance management but called for a stronger focus and more direct relationship with performance management in order to influence resource allocation and choice of activity.[7]

Most of the PSA indicators cover progress towards key targets linked to the Millennium Development Goals. Owing to significant lags between current spending and development outcomes, the value for money indicator is the only PSA indicator that reflects current operational performance. As such, it is a critical measure for DFID.

This study starts by reviewing the coverage and content of the VFM indicator. Next, we examine guidance to staff and internal procedures. Compliance is summarised, followed by a review of trends in performance scores with observations on current operational practice in reporting from a small sample of telephone interviews. The study ends with a summary of key issues and conclusions.

[4] See DFID Public Service Agreement, 2003–2006; and DFID SDA, 2003–2006.
[5] DAC Journal, Volume 2, No. 4. United Kingdom: Development Co-operation Review Main Findings and Recommendations
[6] DFID (2002) Development Effectiveness Report.
[7] NAO (2002) Department for International Development. Performance Management–Helping to Reduce World Poverty.

3. THE VFM INDICATOR

The total volume of aid spent per year by DFID is £2.9 billion (2001/2). Of that amount, £1.5 billion is spent through the bilateral programme; £1.3 billion is spent through multilateral organisations and £88 million is spent on administration (Table 1).

Table 1. DFID External Assistance Programmes (Current Prices, £ millions)

	1998/9		1999/2000		2000/1		2001/2	
Bilateral Aid	1,164	49%	1,324	51%	1,415	50%	1,506	52%
Multilateral Aid	1,131	48%	1,180	46%	1,298	46%	1,315	45%
Administration	66	3%	76	3%	88	3%	88	3%
FAO & UNESCO	**13**	**1%**	**13**	**1%**	**9**	**0%**	**15**	**1%**
TOTAL	**2,374**		**2,593**		**2,810**		**2,924**	

Source: DFID (2002a: 71)

The VFM indicator reports on all project and programme spending through the bilateral programme that are larger than £1m (was £500,000 prior to April 2002[8]), but excluding the first two years of operation. It is calculated from data captured by DFID's PRISM. PRISM is designed to report all project and programme spending, including all of DFID's bilateral spend, plus any assistance to multilateral agencies that falls within the Bilateral Aid Framework.

While PRISM covers all projects and programmes (as extracted from the central management information system (MIS)), the amount of information stored about each project/programme varies considerably. Roll out of PRISM to departments which manage approximately 90% of bilateral spend was achieved in May 2001. [9]

Table 2. Percentage of £1m+ Operational Projects and Commitment[10], from PRISM

	Number of projects	Commitment (£ million)
£1 million (plus) projects	1086	6,295,042
Total number of projects	6775	7,172,588
	16.0%	**87.8%**

Source: PRISM data

Table 2 shows that around 16% of currently operational projects are required to score, being projects or programmes over two years old with over £1 million commitment. This represents some 88% of project financial commitment. Assuming that approved commitment is spread (reasonably) consistently across each financial year, it could be argued that this is equivalent to around 88% of total annual spend, which would represent around £1,323m[11] (45%) of DFID's total expenditure in 2001/2.

[8] Although it is stated as £250,000 in 'Lightening the burden of DFID programme cycle procedures', minutes from Mark Lowcock, Director, Finance and Development Policy, 25 March 2002. This is an error, with £250,000 being used as a common threshold for a number of other project cycle processes at the time, though not scoring.
[9] DFID (2002), Annex 1: Progress against targets in DFID's PSA 2001/2–2003/4, extract from the Autumn Performance Report.
[10] This is approved commitment per project, as opposed to expenditure per year. PRISM stores financial information per project (such as 'approved commitment' and 'expenditure to date'), while other DFID budget systems produce data on an annual basis (i.e. per financial year).
[11] 88% of DFID's bilateral spend (£1,506m) in 2001/2.

As a measure of DFID's bilateral programme,[12] the VFM captures a large proportion of DFID's expenditure. However as this represents only 16% of operations, the VFM excludes a large majority of DFID's operational management (including staff time) as well as potentially important development effectiveness; indeed it is possible that many larger projects involve more capital procurement, whereas many smaller (under £1m) projects encompass some of DFID's 'flagship' or more innovative work, including the policy and influencing agenda.[13]

3.1 Data Capture by PRISM

The completeness and quality of the PRISM data set appears to have improved significantly in recent years, but some anomalies still appear. For example, of the 'approved commitment' figures, a significant proportion were either missing or blank. Out of a total of 31,519 projects stored on PRISM (all years), the approved commitment for 1,044 projects was blank (3.3%) and 4,293 projects were zero (13.6%). This may be a cause of some concern, though further analysis may be able to identify a number of reasonable explanations. These include the fact that the 'approved commitment' is blank because projects are still in the planning stage, or that a large proportion of the data is from older projects (entered retrospectively). Another explanation is that the PRISM data set covers all DFID spending activities, including some miscellaneous programme running costs like medical expenses and training costs. This latter category is given a second level '599' code, and are said to represent the bulk of missing commitment values.

3.2 Measuring Value for Money

The VFM indicator is measured as the percentage of projects rated 1 or 2 on a scale of 5. It is calculated in three separate classes for projects classified by high, medium or low risk status. For example:[14]

$$\frac{\text{Total commitment value of high risk projects approved at Director level and above, scoring 1 or 2}}{\text{Total commitment value of all high risk projects approved at Director level[15] and above (excluding those scoring x)}}$$

The rating scores and risk categories are shown in Box 1.[16]

The simplicity of the measure may conceal some potential technical difficulties. Firstly, as a greater proportion of the bilateral aid programme is transferred to low income countries, it may be reasonable to expect an increase in the commitment value of all high risk projects. In that context, it may become more difficult to achieve scores of 1 or 2 (owing to the greater risk of a project failing)—in other words, a decreasing numerator with an increasing denominator. While this makes the achievement of a sustained increase in success rates more challenging, it also means that the risk separation within the VFM indicator may become a more prominent part of the measurement.

Box 1:Ratings and Risk Categories

[12] Multilateral spend (some 45% of DFID's investment) is not the subject of this paper, even though it represents a massive potential gap in coverage. Indeed work is currently being undertaken into ways of monitoring multilateral expenditure, with some elements being covered within the Service Delivery Agreement (SDA).

[13] It should be noted that elsewhere in the PSA and SDA there are opportunities to capture DFID's innovative and influential work, though the extent to which this captures smaller projects (>£1m) has not been fully explored.

[14] DFID (2002), Technical Note to 2003–2006 PSA, draft 16 (25 September 2002).

In practice many projects of £1m and over are included, even though they are approved by levels lower than that of Director. For example, Heads of

[15] Department can approve projects up to £7.5m.

[16] DFID Office Instructions II: G1 Annex 3

Achievement Rating	Achievement
The following rating scheme should be used to rate the likelihood of achieving outputs and in turn fulfilling the project's purpose. 1 = likely to be completely achieved 2 = likely to be largely achieved 3 = likely to be partially achieved 4 = only likely to be achieved to a very limited extent 5 = unlikely to be realised x = too early to judge the extent of achievement	Projects can be categorised into one of three categories of risk as follows: H = High risk M = Medium risk L = Low risk

Secondly, as projects are not required to score until the end of their first two years, the measure reports on on-going operations in the portfolio rather than new entrants. Furthermore, the scores are designed to be forward-looking (are the outputs and purpose *likely* to be achieved?). Together, these aspects mean the VFM is reporting potential future performance of two-year or older operations.

3.3 The VFM Baseline

The increase in the proportion of DFID's bilateral programme going to low income countries (from 78% to 90%) is largely seen as on-course (Table 3), with both its attainment and measurement falling mostly under the control of DFID's central management.

Table 3. Country Specific Bilateral Aid for Low Income Countries

	1997/8	1998/9	1999/2000	2000/1	2001/2
Spend (£ millions)	535	664	709	886	944
% of Bilateral Aid	65	73	67	76	78

Source: DFID (2002a: 76)

In May 2001, baseline figures for the VFM were calculated based on commitment value, as 24% for high risk projects; 56% for medium risk projects; and 81% for low risk projects.[17] The PSA Technical Note states that the VFM will in future be calculated using commitment value, and the present situation shows an increase above the baseline (except for medium risk projects).

Table 4. Summary of VFM Indicators, Q4 2002

	Number of projects	VFM by number of projects (%)	Commitment (£m)	VFM by Commitment (%)
High risk	30	46	269	35
Medium risk	172	55	1204	49
Low risk99	78	647	84	

Source: Management Report, Q4 2002

[17] DFID (2002), Annex 1: Progress against targets in DFID's Public Service Agreement 2001/2–2003/4, extract from the Autumn Performance Report. It should be noted that the document actually states that the baseline figures are calculated by project numbers, but this an erroneous statement.

3.4 Frequencies of Overall Scores

The current frequency distribution of the scores 1–5 (see Figure 5) highlights the basic characteristics of the rating system, with the majority of projects receiving a score of 2 or 3. The frequency distribution is remarkably consistent, even for pre-2001 years when compliance was much lower and the relatively small dataset shows irregularities in other assessments. Indeed any trend in scoring is not particularly dramatic, with a marginal decline in the proportion of projects receiving a score 2 and an increase in the projects receiving a 3.

Figure 5. Frequencies of Project Scores[18]

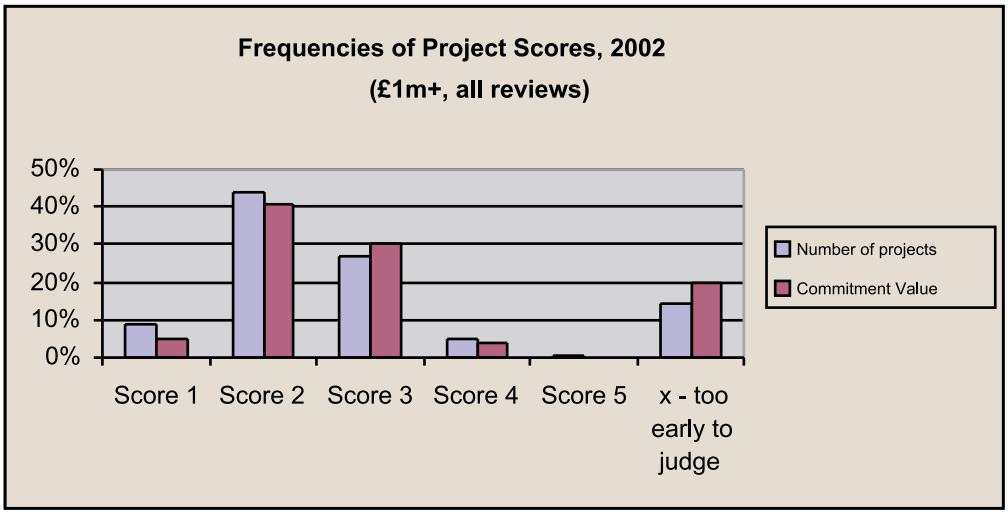


Figure 5. Frequencies of Project Scores

Very few projects receive a score 4 or 5, with the difference between success and failure becoming largely the difference between the interpretation of whether the purpose has been 'largely' or 'partially' achieved (i.e. a score 2 or score 3, respectively). For high risk projects this becomes especially pertinent, as the overall success (of the VFM) ignores 'partial success' (score 3) which might be considered a form of success for highly innovative, ambitious projects in high risk environments. There is also a significant proportion of projects that are rated as 'too early to judge'. These are not included in the VFM indicator, but represent some 14% of all projects over £1m, and 20% of commitment value. This reduces the coverage of the indicator from 16% of projects to 14%, and 70% of commitment value.

[18] Calculated as the number of projects, as a percentage of the total number of project scores (all scores, including 'X') in a year. Repeated for each score from 1 to 5, including 'X'.

Summary of Findings —The VFM Indicator

- The VFM indicator covers a potential 16% of all bilateral projects by number and 88% by commitment value; but after adjusting for projects rated as 'too early to judge' actually captures 14% of all bilateral projects by number and 70% by commitment value.
- Although comprehensive it means that lower value, perhaps very influential or innovative, operations associated with policy development and influencing are excluded.
- A shift of aid to low income countries has the potential to increase the proportion of high risk operations, thus making the target of improving success harder to achieve.
- The indicator reports potential future performance of the two or more year old projects in DFID's portfolio, not current performance of the whole portfolio.
- Scores are clustered among ratings 2 and 3. As the indicator reports on scores 1 and 2, the distinction between success and failure hinges on the interpretation of a project purpose being likely to be 'largely' (2) or 'partially' (3) achieved.

4. GUIDANCE AND INTERNAL PROCEDURES

As mentioned previously, the VFM indicator has two components: a categorisation of riskiness of the operation on a 3-point scale and a score on a 5-point scale of the probability that an operation will achieve its purpose. The VFM reports the percentage of all eligible operations scoring 1 or 2 out of 5, for each risk category high, medium or low (see Box 1).

4.1 Guidance on Assessment of Riskiness of an Operation

The risk status of an operation (high, medium, or low) is the main means of stratifying operations for performance scoring. An assessment of risk is required in the initial project header sheet, in parts 5 and 6 of a project or programme submission, and in the analysis of assumptions in the logical framework. The DFID Office Instructions (OI) do not contain any information about how the risk assessment is to be made, or what distinguishes low, medium and high. Reference is given to Technical Note Number 12, *'The Management of Risk in DFID,s Activities,.*[19] But this document does not provide guidance about the factors to be taken into account when assessing the risk of operations. It is not clear if the risk refers to the environment within which the project will operate, or the nature of the project intervention.

It appears that the risk assessment is made at a single point of time, when the project is designed. There is no requirement to re-assess the risk status during the life of the operation.[20] The absence of guidance means that individual staff, sector teams or Country Offices have a high degree of flexibility about interpreting this classification.

OI Guidance on the treatment of risk:
II:D5 Annex 1
II:D6
Technical Note No. 12

4.2 Guidance on Scoring Operations

Information in OI about the process of scoring has two aspects: the rules about which operations are eligible to be scored and the frequency of that scoring; and the nature of the scoring assessment.

Guidance about eligibility and reporting frequency can be found in OI II:G1. Projects with a commitment value of £1 million or greater are to be scored, smaller projects are exempt. Projects are also exempt from reporting a performance assessment during the first two years of operation. This rule is to reduce the burden of reporting at a stage when many operations are still in a process of gearing up.

OI II:G1 Annexes 1, 2 and 3 contain instructions for monitoring and reporting. Annex 3 contains the PRISM on-going project scoring summary sheet. That sheet instructs the user in the achievement rating and risk category. The information given is shown exactly as it appears in Box 1.

No additional explanation or guidance is given about how the ratings scheme should be interpreted or applied.[21] Additional information is requested on the summary sheet in order to support the rating given. The reporter is asked to provide a narrative justification for both the rating of purpose

[19] In addition to Technical Note Number 12, some guidance is given in the 'Tools for Development' handbook, as well as induction training as part of the 'Programme & Project Cycle Management' course (PPCM), which is shortly to be updated. Further guidance notes issued in August 2003.
[20] It is now a requirement to state the risk level when projects are scored in PRISM. Departments are advised to reassess risk at this point and record the most current risk level, Hugh McGarvey, March 2003, pers. comm.
[21] Guidance notes on these aspects issued in August 2003.

and of outputs; there is space to comment on the extent to which the achievement of project purpose can be attributed to project outputs; and the quality of scoring should be explained in terms of the methodologies used and team composition.

The ratings scheme contains both a set of descriptive phrases and a numerical 5- point scale.[22] There is an extensive literature on the effect of specific wording and scales used for opinion polling, such as measures of satisfaction. Box 2 illustrates some points from a recent paper prepared by MORI for the Cabinet Office.

Box 2

The modifying adverb on verbal scales

A commonly used verbal scale is:

- Very satisfied
- Fairly satisfied
- Neither satisfied nor dissatisfied
- Fairly dissatisfied
- Very dissatisfied

However a simple aggregation of the top two ratings, of 'very' and 'fairly' satisfied responses (as is common practice) will cover a very wide range of service experiences and attitudes. 'Very satisfied' in general represents a positive statement, whereas 'fairly satisfied' results in a less uniform set of interpretations, with researchers' findings that some people feel that it does not even convey a generally favourable statement. The modifying adverb (such as fairly, quite, slightly, etc) has a clear effect on the response elicited.

Numerical rating systems

The assumption that there is equal distance between each point on a numerical scale is just as likely to be incorrect as the assumption that there are equal distances between points on a verbal scale. For example, respondents may interpret the scale with reference to their days at school; on a 1–10 scale, any score below 5 is likely to be regarded as particularly low, with clustering of responses around points 5–8. This makes interpretation of results more complex and less consistent.

Generally the more points used, the more reliable the results, as fewer points on the scale encourages respondents to treat the alternatives as discrete rather than continuous variables. The literature suggests 5–11 points are used.

Source: MORI 2002

The ratings scheme is based on what appears to be simple, non-technical language, with a common-sense aspect about the meaning of completely, largely or partially achieved. As the MORI study shows in Box 2, the wording of the phrase may itself affect the rating. Such an approach may be easy to apply in some instances, where operations have a narrow focus, but more

OI Guidance on rating achievement:
II:G1
II:G1, Annexes 1, 2, 3
II:G2, PRISM

[22] Comparable arrangements used by the World Bank have a 4-point scale, described in Annex 1.

difficult where operations have multiple or diverse components and the assessment involves an implicit weighting. There is no guidance or explanation about how to approach such cases. Neither is any explanation given about linking the flow of information from the implementing agency, the reporting against objectively verifiable indicators (OVIs) set out in the log frame and the project summary score.

4.3 The Monitoring and Reporting System

Project scoring is part of the wider performance management system at DFID. Related elements are:

- OPR—output to purpose reviews, which are the main source of information to support a performance score.
- PCR—project completion reports, written for all projects greater in value than £1m when expenditure has reached at least 95%.
- Evaluation reports—carried out on a very small sample of operations, often in thematic clusters.
- APPR—annual plan and performance review of country programmes. A new review process, not yet effective owing to limitations in the structure and content of programmes and substantial variation in approach among divisions in DFID.

A brief review of the performance reporting system shows that:

- There is no inter-linkage between these instruments, e.g. there is no systematic review of the consistency of annual ratings during implementation with the later PCR.
- There is no link between project or programme rating and the APPR.
- Entry onto PRISM does not trigger a management response. For example, neither a 1 nor a 5 sets in motion any particular response, to seek a confirmation, or to initiate follow-up scrutiny.

Summary of Findings—Guidance and Internal Procedures

- There is no guidance in DFID's OI for the interpretation and use of the risk classification. Staff have discretion in deciding on the type of risk and approach to applying this rating.
- Guidance on the performance ratings is limited to what has to be done, and when. There is no information about how to apply the ratings.
- Both the nature of the wording and the number of points in a ratings scale can have a significant influence on how ratings are used.
- The performance reporting system has no inter-linkages between its constituent elements, which could provide a cross-check of ratings or stimulate follow-up action.

[23] See DER, para 7.12. APPRs are now being replaced by a new system of country assistance planning and review, focused on Country Assistance Plans (CAPs).
[24] Recent examples of the DFID Management Report have included specific reference to following up the performance scores of projects.

5. COMPLIANCE

The compliance rates for project scoring have been reported for the past five Quarterly Management Reports, since Quarter 4 2001 (see Table 6). The number of eligible projects has remained broadly constant within each quarter, while there has been a significant increase in the scoring of projects. This compliance 'catch-up' is particularly dramatic over Q4 2001 to Q1 2002, where the number of projects scored rose from 19% to 52% of eligible projects. The highest compliance rates occur in the Asia region, now exceeding 70% of eligible projects, as well as with the largest increase (from 31% to 78%) from Q4 2001 to Q1 2002. The improvements in Asia came at first from Bangladesh, but now reflect the region as a whole. Compliance rates in Africa, although increasing from a very low starting point, remain relatively low, at less than half of all eligible projects being scored.

Table 6. Number of Eligible Projects for Scoring and Projects Scored 1–5, with Risk

	Threshold £500k					Threshold £1m			
	2001 Q1	2001 Q2	2001 Q3	2001 Q4	2001 Q4[26]	2002 Q1	2002 Q2	2002 Q3	2002 Q4
Africa					**12%**	**39%**	**46%**	**40%**	**36%**
- Eligible projects		435	381	385	305	287	224	225	219
- Scored 1–5, with risk					37	113	104	90	78
Asia					**31%**	**78%**	**74%**	**74%**	**70%**
- Eligible projects		397	361	379	235	209	185	168	161
- Scored 1–5, with risk					72	164	136	124	113
EEWH					**15%**	**40%**	**49%**	**47%**	**46%**
- Eligible projects		179	180	203	198	175	108	137	131
- Scored 1–5, with risk					29	70	53	65	60
Central									
- Eligible projects							160	161	171
- Scored 1–5, with risk							6	8	11
TOTALS					**19%**	**52%**	**57%**	**53%**	**49%**
- Eligible projects		1011	922	967	738	671	517	530	511
- Scored 1–5, with risk					138	347	299	287	262

Source: DFID Management Reports, 2001 to 2002.

Compliance

Summary of Findings—Compliance

- A major 'catch-up' in reporting occurred between Q4 2001 and Q1 2002.
- The highest compliance rates occur in the Asia region (70% of eligible projects).
- Compliance rates in Africa remain relatively low, at less than half of all eligible projects.

6. TRENDS

This section explores some of the recent trends in the scores that are used to make up the VFM indicator, including: the risk categories, regional trends, sectoral trends, and different aid instruments (projects and programmes; SWAps and Budget Support).

6.1 Trends by Risk Categories

As mentioned earlier, there is a concentration of scores 2 and 3, but this is especially so for medium and low risk projects. Relatively few projects score 4, and even fewer score 5, across all risk categories. However a key concern here is that a lot of projects do not state a risk rating, and are marked 'Not Stated' on the PRISM dataset. This is unlikely to be because projects failed to assess risk, but rather that the risk rating has not been stated in a review (when entered on PRISM). This is probably because projects are categorized for risk as a one-off activity during the design stage. It may therefore be the case that, rather than failing to assess risk, Country Office staff do not see a need to enter another risk rating (see Figure 7).[27]

Figure 7. Number of Projects by Scores and Risk Category, 2002

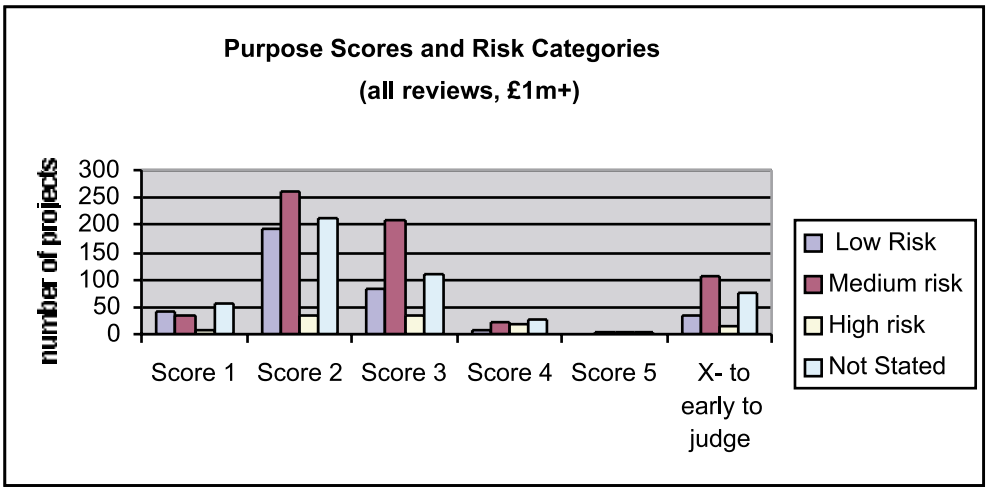
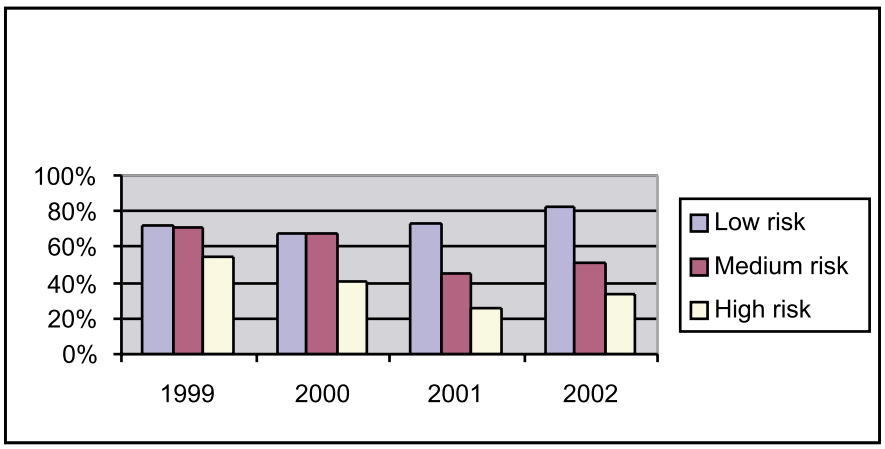


Figure 8 shows a steady increase in successful low risk projects (in terms of total approved commitment); rising from 72% in 1999 to 83% in 2002. Success in medium to high risk projects appears to be falling, with only 51% of medium projects scored 1 and 2 in 2002, and 34% of high risk projects. The percentages for high risk projects suffer from the problem of a low sample size. In 1999 for example, the total number of high risk projects was only seven (excluding those marked 'x–too early to judge'). This helps contribute to an appearance of a high success rate in one year, while in others it drops considerably.

The variation in the patterns of success amongst high, medium and low risk projects presents concerns about the stability of the VFM indicator. However, in the latter years (2001 and 2002) the scores appear to be settling down, conforming more closely to a pattern that might reasonably be expected, and this appears to look even more rational when viewed in terms of commitment (rather than number of projects).

[27] It is now mandatory to input a risk rating when entering a review, and for this financial year 95% of projects scored by regional spending departments have a risk level entered, Hugh McGarvey, March 2003, pers. comm.

Figure 8. Percentage Success Rates by Approved Commitment [28]



6.2 Regional Trends

The number of projects in the DFID portfolio is fairly evenly split between Africa and Asia, especially for £1 million plus projects, at 40% and 41% respectively. However Asia accounts for the largest share of commitment value (some 64%) with projects on the whole being significantly larger (Figure 9).

Figure 9. Profile of Project Portfolio by Region (Q4, 2002)

Table 3. Country Specific Bilateral Aid for Low Income Countries

	Africa	Asia	EMAD
Number of projects	97	113	68
£1m+ projects only (%)[29]	40	41	30
Total commitment value	£540m	£1,280m	£180m
Average commitment value	£5.6m	£11.3m	£2.6m

Source: Management Report Q4, 2002

The proportion of projects scoring 1 and 2 varies between the Africa and Asia regions. For scores 1, the variation is most pronounced (Figure 10) where there has been huge fluctuation, particularly in years 1999 and 2000. This however is largely a feature of low compliance and thus a small sample size in those years. In 2002, the regions are more closely matched for both scores 1 and 2 (see Figure 11).

[28] Calculated as approved commitment (£'000) for projects scoring 1 and 2 (low risk), as a percentage of total approved commitment (low risk). Repeated for medium and high risk categories.

[29] This is an approximate measure of the number of £1m+ projects, based on the number of reviews.

Figure 10. Percentage[30] of Score 1 only, for Africa and Asia

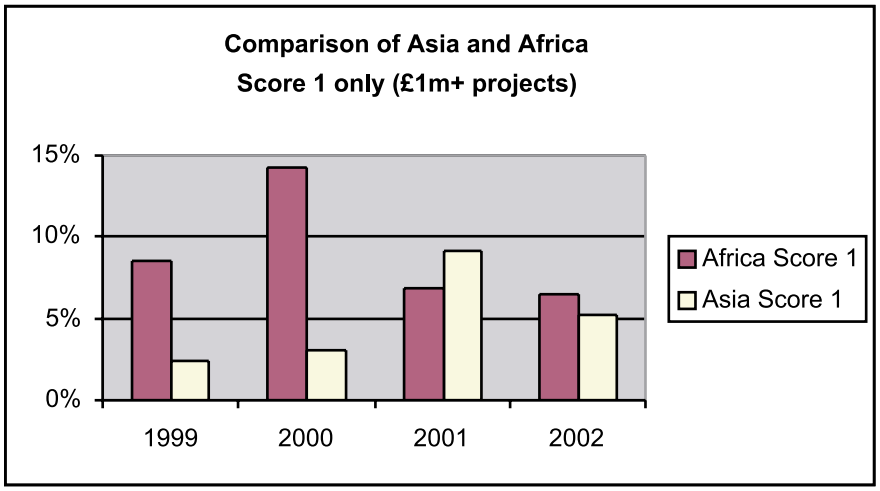
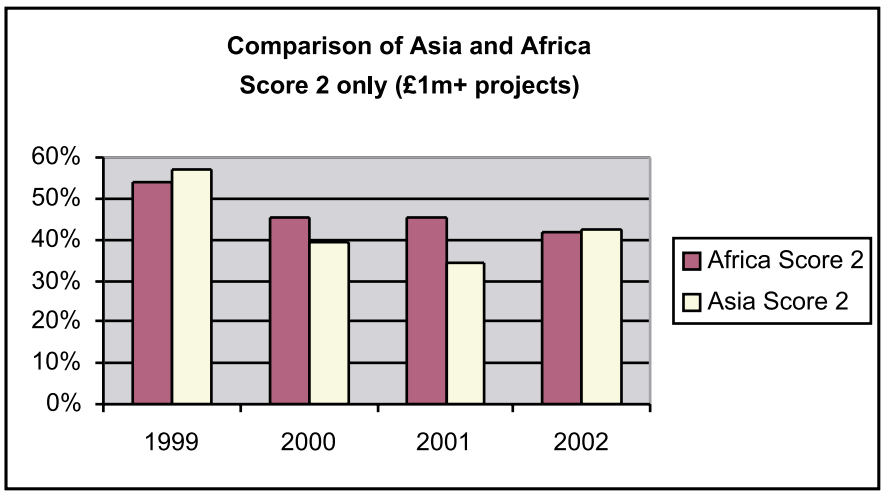


Figure 11. Percentage of Score 2 only, for Africa and Asia



It is possible that regional differences can be attributed to the pattern of scoring by particular DFID departments. However, owing to the relatively small sample size caused by disaggregating in this way, it becomes difficult to conclusively isolate such patterns (Table 12). For Africa and Asia, tentative conclusions might suggest that DFID Eastern Africa, India, Nepal and the Western Asia departments rate purposes as predominantly score 2, whereas DFID Southern Africa and Bangladesh rate predominantly as a score 3. A high proportion of 'x–too early to judge' ratings are given by DFID Central Africa and the Eastern Asia And Pacific Department. The small numbers involved however mean that the scoring of just one project can cause significant anomalies to occur. For an example of score clustering, note that all of the Middle East and North Africa Department projects scored only a 2 or a 3.

[30] Calculated as the number of score 1s as a percentage of the total for scores 1–5, repeated per region.

Table 12. Distribution[31] of Purpose Ratings (£1m+ projects) by DFID Department, 2002

	No.	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	x (%)	Grand Total (%)
AFRICA								
DFID Eastern Africa	31	10	68	16	3	0	3	100
DFID Central Africa	28	7	25	11	18	0	39	100
DFID Southern Africa	24	0	38	50	8	0	4	100
West And North Africa Department	8	0	38	50	13	0	0	100
Africa Great Lakes and Horn Dept	4	25	50	0	0	0	25	100
Average (Africa)	**95**	**6**	**44**	**25**	**9**	**0**	**15**	**100**
ASIA								
Western Asia Department	24	0	58	25	0	0	17	100
DFID India	23	4	61	22	4	4	4	100
DFID South East Asia	14	18	55	9	0	0	18	100
DFID Bangladesh	13	8	38	50	4	0	0	100
Eastern Asia & Pacific Department[32]	12	0	36	7	0	0	57	100
DFID Nepal	11	0	62	31	8	0	0	100
Average (Asia)	**97**	**5**	**51**	**27**	**3**	**1**	**13**	**100**
EMAD								
DFID Caribbean	20	33	50	17	0	0	0	100
Central & South Eastern Europe Dept	12	13	13	13	25	0	38	100
Middle East and North Africa Dept	10	0	70	30	0	0	0	100
Latin America Department	9	33	33	22	0	0	11	100
Overseas Territories Department	9	22	44	33	0	0	0	100
Eastern Europe And Central Asia Dept	8	5	55	35	0	0	5	100
Average (EMAD)	**68**	**16**	**47**	**26**	**3**	**0**	**7**	**100**
OTHER[33]								
Average (Other)	**9**	**22**	**33**	**22**	**0**	**0**	**22**	**100**

The regional distribution of risks in Table 13, shows that there are wide discrepancies between Africa and Asia. Asia has a similar number of high risk projects as Africa, but they are double the commitment value. Medium risk projects are similar in number between Africa and Asia and higher in value in Asia. But for low risk, Asia has double the number of projects and four times the commitment value. The greater share of low risk in Asia is more intuitive than the higher level of high risk. Africa is more commonly considered to be a higher risk environment for undertaking projects. But as noted earlier, it is not clear how the risk rating is being used—to rate the nature of the project intervention, or the environment within which it is operating.

The difference in size of the portfolio between Asia and Africa means that the generally larger Asia projects (in monetary terms) have a greater impact on the overall VFM indicator, as it is measured by commitment value. Comparison of the VFM indicator for Africa and Asia in Figure 14 shows similar ratings (by value) for low and medium risk project, but a marked difference for high risk. The

[31] This is the calculated as the number of scores 1 as a percentage of score 1–x, repeated for each score.
[32] It should be noted that since this analysis in January 2003, Eastern Asia and Pacific Department (EAPD) has now achieved a much higher level of compliance and the percentage scoring a 'X' has fallen to 22%.
[33] Includes Environment Policy Department, Health And Population Department, Rural Livelihoods Department and Social Development Department.

VFM for high risk projects in Africa is 82%, yet for high risk projects in Asia it is 11%[34]. DFID's 'high risk' bilateral programme would appear to be more at risk in Asia than Africa.

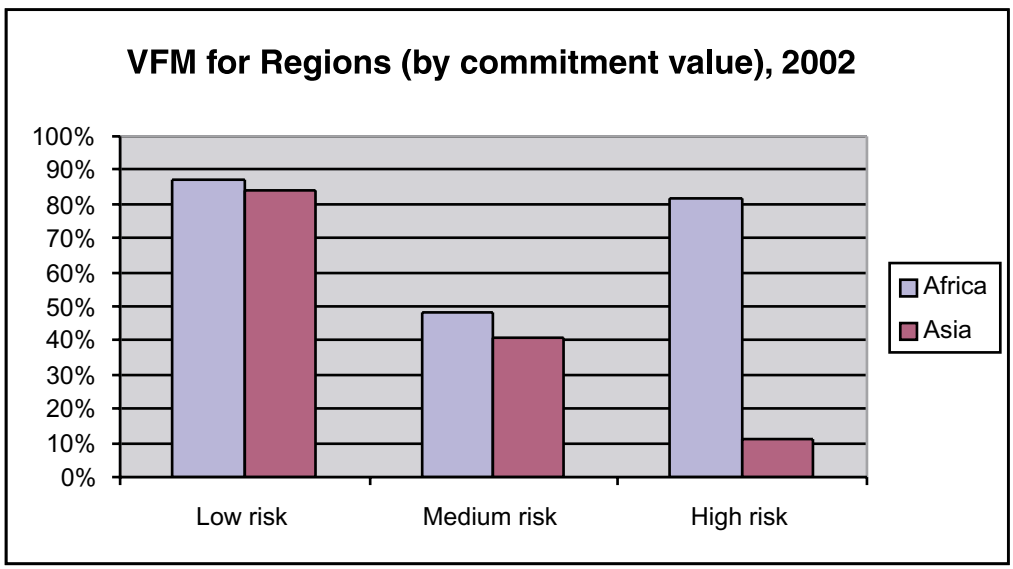
Table 13. Regions Classified by Risk Category (High, Medium, Low) for 2002

	Africa	Asia	EMAD	Other	TOTALS
Low risk					
number of projects	24	51	22	2	**99**
percentage of projects	*24%*	*52%*	*22%*	*2%*	***100%***
commitment value	£108m	£490m	£43m	£4m	**£645m**
percentage of value	*17%*	*76%*	*7%*	*1%*	***100%***
Medium risk					
number of projects	62	61	42	7	**172**
percentage of projects	*36%*	*35%*	*24%*	*4%*	***100%***
commitment value	£357m	£611m	£130m	£104m	**£1,202m**
percentage of value	*30%*	*51%*	*11%*	*9%*	***100%***
High risk					
number of projects	11	12	4	3	**30**
percentage of projects	*37%*	*40%*	*13%*	*10%*	***100%***
commitment value	£75m	£179m	£7m	£5m	**£266m**
percentage of value	*28%*	*67%*	*3%*	*2%*	***100%***

Source: Management Report Q4, 2002

It is argued that it is perhaps too early to assess these wide discrepancies, although they are so different that they may reflect a problem with the structure of the scoring system[35]. In particular, with limited guidance, the risk assessment is unclear. This makes it difficult to compare between regions, and possibly between countries.

Figure 14. Scores 1 and 2 by Region (Commitment Value), 2002



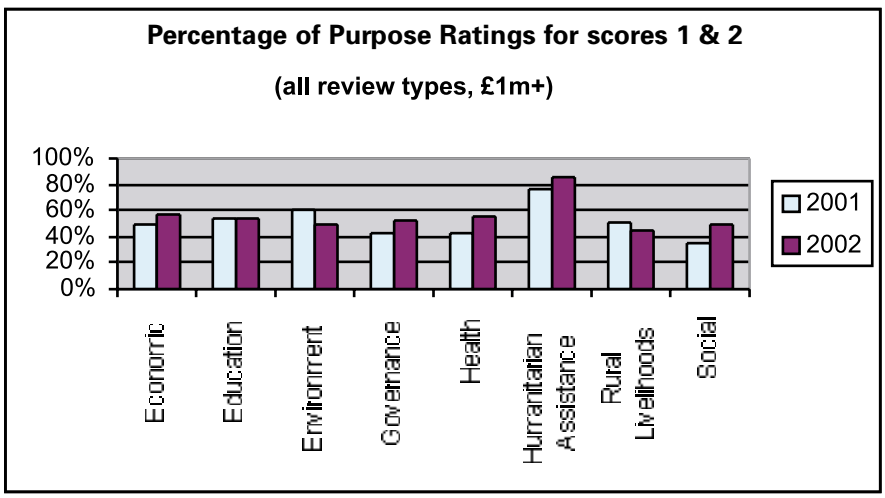
[34] Management Report, Q4 2002.

[35] Page 13, Management Report, Q4 2002.

6.3 Trends by Sector

There is nothing particularly striking about the percentage of scores 1 and 2 given by each sector, except that Humanitarian Assistance scores extremely highly (over 80% in 2002). 'Humanitarian assistance' covers relief aid and reconstruction projects undertaken by bilateral programmes (such as DFID India following an earthquake). It includes only a very small percentage of emergency relief undertaken by the Conflict and Humanitarian Affairs Department (CHAD). It is possible that the success of such work (e.g. the distribution of agricultural seeds) is easier to judge objectively as it is geared more to delivery of outputs than to a developmental purpose. Other sectors score less well, and in particular the 'social' sector where projects receive a much lower proportion of scores 1 and 2. It is however difficult to know what constitutes 'social' as the category is broad and selected by the department with no specific criteria.

Figure 15. Scores 1 and 2 by Sector, 2002



6.4 Instruments (SWAps, DBS)

Overall, very few SWAps have been scored, though this is primarily due to the small numbers involved, rather than as a consequence of non-compliance. At the time of writing, there were ten currently operational SWAps (over £1m) scored, which represents some 83.3% of all eligible SWAps. Of those which have been scored, most appeared to have been scored cautiously, receiving scores 2, 3 or 'x–too early to judge'. For currently operational DBS there are only five that have been scored and entered onto PRISM, some 85.7% of all eligible DBS. With a rising trend of SWAp and DBS operations however, the scoring and performance assessment of SWAps/DBS is potentially significant for future PSAs. This is especially so as SWAps and DBS represent an important slice of DFID expenditure at £323m and £341m respectively.

Summary of Findings—Trends

- The pattern of scoring by risk has settled down during 2001 and 2002 to an intuitively plausible distribution overall, with best performance for low risk, less for medium and least for high.
- Similarly, for comparison between the major regions Asia and Africa, results in 2001 and 2002 show fewer anomalous fluctuations than earlier years.
- The relatively higher commitment value of projects in Asia means that performance in that region has a dominant effect on the overall indicator.
- Africa has a higher number of projects being scored 1, than has Asia.
- Relatively small numbers of projects make comparison at country or sub-regional level difficult. Apparent anomalies may in fact be spurious.
- Analysis of regional performance by risk reveals a similar number but much higher value of high risk projects in Asia compared with Africa. High risk projects in Africa have much higher VFM scores than in Asia, suggesting that DFID's high risk bilateral programme is more at risk in Asia.
- The current inconsistent treatment of risk renders regional comparisons unreliable.
- No significant trends emerge for scoring by sector; very few SWAps or DBS programmes have been scored.

7. EVIDENCE AND PRACTICE TO SUPPORT SCORES

Purpose scores are inputted onto PRISM from a range of monitoring mechanisms (Table 15). Most commonly, purpose scores are a result of an OPR or a PCR. Indeed, over the past two years, OPRs and PCRs represent some 66 to 70% of all scores entered onto PRISM. However, other review types are cited as an important part of the scoring data. These include the Snapshot Annual, Progress and Monitoring reports—though these categories are selected by the data inputer and do not necessary reflect institutionalised processes.

Table 16. Purpose Ratings by Review Type (£1m+ projects)

	2002	
	Scored	**%**
PCR	99	37
OPR	77	29
Snapshot Annual	36	13
Progress	23	9
Monitor	21	8
Unknown/blank	12	4
Other[36]	1	0
Grand Total	**269**	**100**

7.1 Documentary Evidence

According to records in PRISM only around one third of OPRs and PCRs have supporting documentation loaded onto PRISM. This does not mean that the evidence does not exist, but that it is not available to users. PRISM requires scores to be loaded onto the system by using either a data entry screen or a MS Word template from which the system is able to extract the information. Supporting documentation, such as the full OPR report are then emailed to the PRISM team for uploading onto the system. Table 17 shows the number of OPRs and PCRs supported with documentary evidence—either documents uploaded onto the system as 'OPR' or 'PCR', or reports labelled 'Consultancy Report', 'Monitoring Report' or 'Other Report'. On average, around 30% of OPRs or PCRs are supported with their respective documents, plus an additional 17–22% accounted by other documentation. These figures suggest that about half of supporting evidence is loaded onto the system.

The system is intended to ensure that a minimum amount of data should be recorded about all scores, including: the purpose and output justifications, purpose attribution and quality of scoring. In a comparison of the OPR scores for the last quarter of 2001 and 2002, there is an increase in the amount of information completed between the two periods (Table 18). However there remains a significant proportion of OPR scores contained on PRISM which do not have a purpose justification, outputs justification or purpose attribution. This is particularly the case for purpose attribution, which should contain information about the extent to which achievement of purpose can be claimed to arise from the effects of the project. Again this is not necessarily because the information does not exist, but it has not been entered onto PRISM.

[36] Includes 'PCR Exempt', 'Snapshot Quarterly' and 'Inception'.

Table 17. Documentary Support for Scored OPRs and PCRs (£1m+ projects)

	2001		2002	
	Number	%	Number	%
OPRs supported with:				
An OPR document	8	8	23	30
Other Reports[37]	33	33	13	17
Remaining	59	59	41	53
Sub-total	***100***	***100***	***77***	***100***
PCRs supported with:				
A PCR document	43	58	29	38
Other Reports	6	8	17	22
Remaining	25	34	53	40
Sub-total	***74***	***100***	***99***	***100***
Both OPRs & PCRs				
OPR/PCR documents	51	29	52	30
Other Reports	39	22	30	17
Remaining	84	49	94	53
Total	***174***	***100***	***176***	***100***

Table 18. Evidence to support OPR Scores (£1m+ projects)

	2001 (n = 108)		2002 (n = 72)	
	Number	%	Number	%
Purpose Justification				
Yes	74	69	60	83
No data	34	31	12	17
Output Justification				
Yes	56	52	55	76
No data	52	48	17	24
Purpose Attribution				
Yes	57	53	45	63
No data	51	47	27	38

7.2 Review Methodologies

As part of completing the scoring data entry screen or template within PRISM, there is a section marked 'Quality of Scoring'. Completion is optional. The information in this box varies considerably from simple statements like 'satisfactory' to complete explanations of methodology. Many others comment on the monitoring system used to collect data, lessons learnt so far, and provide recommendations for future action. The free format to the information means it cannot easily be analysed for the extent of stakeholder involvement, the use of external consultants and the methods used to conduct the review (desk study, surveys, field visits).

A more detailed review of 30 OPR documents selected randomly from PRISM (see Annex 2) looked at the process of the review, the composition of the review team and the review methodology. Three OPRs (11%) were clearly done by DFID staff only; eleven (39%) were conducted by mixed teams including DFID staff; two (7%) were by external consultants only; and twelve (43%) had insufficient

[37] Includes documents called, 'Consultancy Report', 'Monitoring Report', and 'Other Report'.

information to judge. The mixed teams often included staff from partner donors, staff from the implementing organisation, and consultants. There was little evidence of involvement by other stakeholders.

Five reviews (18%), contained some mention of the review procedure, though often brief. Most were carried out through individual or group key informant interviews, only two mentioned a specific participatory process or event. More comprehensive information was available in the full reports than the proforma documents.

7.3 Quality of Supporting Evidence

The quality of the purpose rating depends on the information used to support that rating, which should be drawn from objectively verifiable indicators of project purpose in the logframe. In the assessment detailed in Annex 2, inspection of the purpose indicators showed that while four out of five projects had indicators with a clearly specified qualitative dimension, only around half the projects had quantified or time-bound indicators (Table 2)[38]. Only one review made reference to baseline data.

Table 19. Quality of Purpose OVIs—Percentage of Projects with QQT Indicators

Quantity	50%
Quality	82%
Timing	57%

Space is provided in the proforma for an explanation of the justification for the selected rating. Some 68% of these were judged to be written in an objective way, drawing on factual information about both the purpose and output indicators, plus other salient information. A significant minority (29%) were more subjective in nature, failing to draw on factual information. One review document did not provide any justification text.

The ratings given for outputs were also assessed. The presentation of this information differs markedly among the documents. Some reviewers rate output delivery overall; others rate each output; and some rate each output indicator. A few documents provide both separate and combined ratings for the outputs. In view of the large number of output indicators, no assessment of the QQT quality of these ratings was made. But the narrative description for each output was reviewed. Some 71% were judged to be objectively written; 23% were more subjective in style and a small proportion, 6%, had no description. The styles used ranged from brief statements echoing the ratings categories: 'delivery likely to be largely achieved'; through a verbal restatement of the indicator; to a contextual and analytical narrative.

The final assessment was a judgement of the extent to which the material in the document supports the given purpose rating. This was not an assessment of whether the rating was correct or not, but only of the coherence of the information. Some 64% were assessed as being in support. In nine of the 28 cases (32%) the reviewer considered that the rating was not adequately supported. Examples include:

[38] A characteristic of a good indicator is that it is structured with a Quantitative measure; has a detailed statement of the nature of the measure Quality; and has statements of when target values will be attained–Timing. QQT.

- not giving a rating because a livelihoods monitoring system was not in place, despite good information at output level and an evident sound understanding of how target beneficiaries were responding;

- a review where the supportive narrative was restricted to the rating phrases (likely to be partially achieved etc.);

- a narrative that said achievement of purpose was unlikely yet a score of 'x' was given owing to the timing of the review; and

- instances where purpose indicators were rated equally as 2s and 3s yet the overall rating was given as 2.

7.4 The Review Process

In order to better understand the processes that underpin project scoring, a number of telephone interviews were conducted with DFID Country Offices. Responses were drawn from DFID Bangladesh, Brazil, India and Malawi including Advisors covering Health, Engineering, Economics, Social Development and Rural Livelihoods. The sample size was fairly small, but it still provides an insightful (though tentative) observation of the reviews undertaken by Country Offices.

In general, the review process between Country Offices varies considerably, but the approach taken appears to be mostly thorough and methodical. In part this stems from the need of an office to be able to measure performance and take appropriate management decisions. Furthermore, some Country Offices are taking innovative steps (such as with SWAps), though experience sharing and lesson learning is limited across (and sometimes within) country programmes. In DFID Malawi, the Health sector undertook a new approach to OPRs by covering three related projects in one review, while in DFID Bangladesh Rural Livelihood OPRs are in the process of being contracted out under a single consultancy contract. For SWAps and Budget Support, DFID Offices are still grappling with the 'project-ised' approach of OPRs—an approach that does not fit comfortably with sector-wide programmes.

The review process tends to be tailored towards the nature of the project and the needs of the Country Office. The general approach is to establish review teams, often in conjunction with the partner (or implementing) agency. In some cases there has been a move towards getting partners to 'own' and conduct the review process themselves. This has been particularly important for SWAps where there may be several donor partners involved in a review, though other examples include the use of reviews as a means of institutionalising the project within government. External consultants are also used in many reviews, though this does tend to vary between sectors and projects. In addition to external (or independent) consultants, many Country Offices call upon the services of DFID Advisors in specialist areas such as Rural Livelihoods, Social Development, Economics, etc.

Stakeholder consultation is seen as an important part of the review process, with field visits and beneficiary consultations also taking place for particular projects. DFID India for example frequently uses field visits as part of the Annual Review process, and this will typically involve speaking to beneficiary groups.

Reviews can sometimes be undertaken as a separate mechanism to the internal requirements of DFID—with the final report (the review) being used to complete OPR proformas, project scoring, etc. after the event. For example, Country Offices may set out to have the review 'owned' by the implementing partners or at least jointly shared with Government staff. DFID staff then use the report/findings to complete internal procedures. On other occasions, external consultants will be former DFID Advisors who are able to deal with internal requirements.

Project scoring is mostly seen as a central need, being of limited use for Country Office management purposes. Indeed Country Offices use a wide variety of ways for measuring performance, and tend to measure in terms of indicators rather than overall scoring. For example in DFID Malawi, SWAps may be measured using a set of process and impact indicators developed jointly with Government and other partners. Also in Malawi, Budget Support is being increasingly linked to key input indicators (with limited outcome indicators), as the first priority is to ensure money is being spent in the key poverty areas. Within this context however, measuring value for money per se is very problematic, especially with the move towards sector-wide and budgetary support. In DFID Malawi (Economics) they are developing a 'poverty-linked expenditure' approach for their budget support programme, where value for money is linked to key poverty reduction areas. Staff in DFID Brazil have developed a matrix of outcomes and indicators showing how they intend to contribute to the VFM in the PSA.

The quality of the review and scoring is usually only checked by the relevant Project Officer, and rarely by the Head of the Country Office. Sometimes the Head of the Country Office will get involved when there are particular issues, or if they have a personal or professional interest in the project. Institutional pressure is difficult to assess, but there does not appear to be any particular pressure against the scoring of 4s or 5s. While it is rare for a project to receive a score 5, it has been suggested that this may be partially due to mitigating action taken earlier in failing projects.

Risk analysis is usually only done at the project design stage, and few Country Offices re-appraise the risk analysis once the project is up and running. In general the risk analysis is seen as a one-off event. Risk may be assessed by the DFID Project Officer only, or sometimes jointly with an external consultant, while for others it is a wholly DFID staffed exercise. The composition of the teams and methodologies for assessing risk vary considerably—and with limited guidance and tools, risk assessments can be rather unstructured. In terms of quality checks, the risk categorisation and analysis (risk matrix) is almost always checked (and sometimes questioned) by the Country Office Head. This seems to be the case across all Country Offices interviewed—and is largely because the Project Memorandum is required to go via the Head of the Country Office.

Summary of Findings—Evidence and Practice to Support Scores

- Some 66% of reviews are reported as arising from an OPR or a PCR. A further 13% are from a 'Snapshot Annual Review'.
- Provision exists for a supporting document to be stored so that it is accessible from PRISM. Of the OPR/PCR reviews, around 50% have documents loaded onto the system.
- Many fields on the PRISM data entry form are not consistently completed. The trend has improved from 2001 to 2002, but some 20 to 40% do not give justification for the purpose score, output score and purpose attribution.
- Analysis of a sample of supporting documents showed skimpy treatment of review team composition and methodology, such that the independence and rigour of the review cannot easily be determined.
- Most review documents appear to have been written in an objective style, with reference to indicators and evidence. Only half the purpose OVIs were quantified and time bound. Coherent support for the ratings was judged to be present in two thirds of the reviews.
- Country Offices are being innovative in their approach to monitoring performance and to completing the performance rating. Approaches vary widely and there is little learning between countries.

8. CONCLUSIONS AND KEY ISSUES

8.1 What Does the VFM Measure?

- The VFM indicator is a forward-looking self-assessment of the likely future performance of a subset of the bilateral portfolio.
- All projects with a commitment value of £1 million that have been in operation for two years or more are eligible. The indicator captures about 70% of the bilateral commitment.
- Owing to the large number of small projects, this equates to 14% of projects by number.
- The higher average value of projects in Asia means the overall indicator is heavily influenced by performance in that region.
- Success is defined as the proportion of projects scoring 1 and 2 out of a 5-point scale. Around 70% of projects score 2 or 3. The defining feature for success therefore, is the separation between 2 and 3, whether a project purpose is likely to be largely or partially achieved.

8.2 Are the Measurements Consistent Across the Portfolio?

- Small numbers of projects within specific countries or sub-regions makes disaggregating below regions unreliable, but there appear to be distinctly different scoring distributions between sub-regional units (Table 12).
- There appears to be a major inconsistency in the classification of risk and scoring of performance according to risk status.
- It is not clear if the £1 million criterion excludes the same types of projects across all regions.
- There are too few SWAps and DBS being scored to be able to assess trends.
- On balance, there is some settling down with increased compliance, but there are also significant inconsistencies in the scoring and risk categorization. These call into question the extent to which it is reasonable for the VFM to aggregate across DFID departments, regions, sectors and instruments.

8.3 Are the Measurements Supported with Evidence?

- Examples of supporting documents show that the majority of these attempt to use factual information to support the ratings.
- But PRISM captures a relatively small proportion of supporting evidence, as only around a third of PCR/OPR documents are uploaded onto the system; a sizeable proportion of the Purpose,

Output and Attribution justifications are not completed; and, the Quality of Scoring is completed in a huge variety of ways.
- Supporting evidence may exist but it is not adequately accessible though PRISM.

8.4 What is the Quality of this Evidence?

- Review of a small sample of documents show some reviews provide evidence that is thorough and objective.

- But it has not been possible to determine the overall quality from this brief review. Narrative statements attempt to draw on indicators but many of these lack a sound structure and many purpose statements in fact reflect delivery of outputs. The wide variations in style and content make comparison difficult.

- Most documents do not make clear how reviews were conducted or the extent to which they are independent of DFID.

8.5 Does the DFID System Effectively Support Quality and Consistency?

- The current guidance in Office Instructions does not provide adequate support about applying the risk and performance ratings.

- The various elements of the performance reporting system operate independently with no provision for cross-checks or reviews of consistency.

- There are no arrangements or guidance about the balance between self-assessment and independence in project reviews and ratings.

- In view of the importance of the PSA targets, a better understanding is needed about how staff respond to the current rating phrases and numbers.

- Despite the decentralised nature of DFID's operations there is no guidance about the use of the system for country, sub-region or sectoral management.

- Innovative approaches are being developed to tackle new instruments such as SWAps, but there is no mechanism to share these across DFID.

8.6 Ways Forward

This study shows that there are a number of limitations to the VFM, particularly in the areas of compliance, consistency of approach and content. The possible remedies are numerous but could include management pressure to increase compliance, better guidance and benchmarking to ensure consistency[39]. However it is difficult to make clear and confident recommendations based on a fairly light review (i.e. a desk study). This is particularly so because the main findings highlight significant concerns over the robustness of the VFM, some of which require better understanding. These include the need to capture the full range of country practice, as well as possible lesson learning on approaches being developed. There are also concerns over how robust the system is to be able to cope with a changing portfolio and moves towards SWAps and Direct Budget Support. And finally, though the system should have value at the level of the reporting unit, there are apparent contradictions in aggregate. Indeed it is perhaps this that offers the best way forward: at the country or regional level a huge potential exists for making project scoring (and risk rating) more relevant to managing the performance of DFID Country Offices, with links to the Country Assistance Plans (CAP).

[39] The primary purpose of this paper is to examine the consistency and robustness of the VFM indicator, and not to extensively set out the actions to be taken forward by DFID. During a review meeting of the draft paper however, suggestions were put forward to improve consistency. These included using statistical (normalization) techniques, sampling projects that fall below the £1m threshold, and running an independent check of say 20% of the portfolio. In addition, the performance of projects (i.e. scoring and risk ratings) could be linked to management responses. This may (or may not) include implications for resource allocation, where care needs to be taken to avoid perverse incentives.

ANNEX 1: SUMMARY OF THE WORLD BANK SYSTEM

The World Bank has a comprehensive system of measures for portfolio performance. Table 1 summarises the main elements, following a project cycle sequence.

Table 1

Process/Event	Description and Comment
Project concept review	Discussion of early draft project proposal, opportunity to rethink design.
Quality enhancement review (QER)	Voluntary review with diverse specialists to examine design issues.
PAD review	Review of a new Project Appraisal Document by the regional department.
Board approval	Approval by the Bank's Board.
QAG QAE	Quality at entry assessment by the Quality Assurance Group. Assessment of a sample of projects (approximately 50/200) each year, main emphasis on Bank processes, too late to influence design.
PSR	Twice-yearly mandatory assessment on a Project Status Report, with financial information, ratings of implementation, development objectives and risk status; plus narrative to support ratings and quote performance indicators. Latter not always present.
QAG QSR	Quality of supervision assessment on a sample of projects.
MTR	Mid-term review: the responsibility of the borrower and not mandatory though majority of projects do.
ICR	Implementation completion report at the end of the project, written initially by the borrower then by the lending department. Includes detailed ratings of many aspects of project performance.
ICR Review	Validation of the ICR ratings by the Banks Operational Evaluation department. Done on all projects; analysis used to compare the eventual discrepancy between OED and ICR ratings—the so-called 'disconnect'.
Performance audits and Impact evaluation	Undertaken ex-post on a sample of projects, often by thematic area.
ARPP ARDE	Two major annual reports. The Annual Review of Portfolio Performance is based on a statistical analysis of the current portfolio and performance measures derived from the PSR and QAG. The Annual Review of Development Effectiveness is produced by OED from evaluation results. There is some overlap between the documents.

Relevant issues

- The PSR is the nearest equivalent to DFID's project scoring. The approaches have much in common but the Bank's form has a wider range of aspects that are rated and draws on supervision missions led by Bank staff rather than any formal reviews such as an OPR. Performance ratings are mostly on a 4-point scale: highly satisfactory; satisfactory; unsatisfactory; highly unsatisfactory.

- Results from the ratings, together with time-based measures (such as date of effectiveness after Board Approval), expenditure statistics, and wider country economic management are used in secondary analysis to identify 'projects at risk' for management attention. This was introduced in part to counter a tendency for optimistic rating in the PSR. A project can be satisfactory but still assessed to be at risk.

- The work of the Quality Assurance Group has been a major factor in recent years in focusing attention on the quality of the portfolio. QAG reviews are independent, but conducted by Bank staff on temporary assignment (a feature generally liked by staff being reviewed) and relatively expensive of staff time with a panel of three or four staff spending up to three or four days on each project. Owing to their timing, they are really quality assessment rather than quality assurance. They are credited with stimulating a quality response in the regional department to ensure new projects are 'QAG-proof'.

ANNEX 2: REVIEW OF OPR DOCUMENTATION IN SUPPORT OF PRISM VFM SCORES

In order to assess the extent to which VFM scores are underpinned by a formal review, we examined a sample of review documents. The sample consisted of 30 reviews. These were drawn at random from a PRISM-generated list of OPR documentation for the Africa and Asia regions. Fifteen documents were sampled in each region. Two were found to contain text only with no rating information and were dropped from the assessment. The remaining 28 were divided equally between Africa and Asia. The dates of reviews spanned the period from June 1997 to December 2002, with 24 in 2000, 2001 and 2002. Sector representation was also wide, ranging from health and education to rural livelihoods, water and infrastructure. One emergency operation was included. Eight of the documents were found to be whole or partial OPR reports; the remainder were either the OPR or PRISM summary review proforma only.

Preview process

The first part of the assessment looked at the process of the review: the composition of the review team and the review methodology. Very little information was evident about this aspect. Three reviews (11%) were clearly done by DFID staff only; eleven (39%) were conducted by mixed teams including DFID staff; two (7%) were by external consultants only; and twelve (43%) had insufficient information to judge. The mixed teams often included staff from partner donors, staff from the implementing organisation, and consultants. There was little evidence of involvement by other stakeholders.

Five reviews (18%), contained some mention of the review procedure, though often brief. Most were carried out through individual or group key informant interviews, only two mentioned a specific participatory process or event.

More comprehensive information was available in the full reports than the proforma documents.

Achievement of project purpose

Table 1 shows the distribution of ratings of project purpose. The concentration of ratings 1, 2 and 3 mirrors the overall results from PRISM. The sample appears therefore to be a fair reflection of the population of rated projects.

Table 1: Distribution of Purpose Rating

1	2	3	4	5	x	Blank
7%	46%	21%	7%	0%	14%	4%

The ability to make an objective assessment of progress depends in part on the existence of good indicators at the purpose level. Key characteristics of indicators is that they should be QQT:

- Q-quantity–have a quantified measure of the extent of performance that is expected, e.g. 200,000 children.

- Q-quality—have a clearly specific nature of the indicator, e.g. 200,000 children under five-years-old living in Dedza rural districts.

- T-timing—have a defined period in which the target will be reached, e.g. 200,000 children under five-years-old living in Dedza rural districts will be immunized by June 2003.

Inspection of the purpose indicators showed that while the majority of projects had indicators with a clearly specified qualitative dimension, around half the projects had quantified or time-bound indicators (Table 2). Only one review made reference to baseline data.

Table 2: Quality of Purpose OVIs—Percentage of Projects with QQT Indicators

Space is provided in the proforma for an explanation of the justification for the selected rating. Some

Quantity	50%
Quality	82%
Timing	57%

68% of these were judged to be written in an objective way, drawing on factual information about both the purpose and output indicators, plus other salient information. A significant minority (29%) were more subjective in nature, failing to draw on factual information. One review document did not provide any justification text.

Ratings of outputs

The ratings given for outputs were also assessed. The presentation of this information differs markedly among the documents. Some reviewers rate output delivery overall; others rate each output; and some rate each output indicator. A few documents provide both separate and combined ratings for the outputs. Table 3 shows the percentage distribution of ratings using data for each output where available, or the best available (indicator or aggregate) where not.

Table 3: Distribution of Outputs Ratings

The concentration of ratings among score 1, 2 and 3 is similar to ratings of purpose but with more outputs rated 1. Minimal use is made of ratings 4 and 5.

1	2	3	4	5	x	Blank
27%	34%	14%	1%	0%	17%	7%

In view of the large number of output indicators, no assessment of the QQT quality of these ratings was made. But the narrative description for each output was reviewed. Some 71% were judged to be objectively written; 23% were more subjective in style and a small proportion, 6%, had no description. The styles used ranged from brief statements echoing the ratings categories: delivery likely to be largely achieved; through a verbal restatement of the indicator; to a contextual and judgemental narrative.

How convincing are the OPR proformas?

The final assessment was a judgement of the extent to which the material in the document supports the given purpose rating. This was not an assessment of whether the rating was correct or not, but only of the coherence of the information. Some 64% were assessed as being in support. In nine of the 28 cases (32%) the reviewer considered that the rating was not adequately supported. Examples include not giving a rating because a livelihoods monitoring system was not in place, despite good information at output level and an evident sound understanding of how target beneficiaries were responding; a review where the supportive narrative was restricted to the rating phrases (likely to be partially achieved etc.); a narrative that said achievement of purpose was unlikely yet a score of 'x' was given owing to the timing of the review; and instances where purpose indicators were rated equally as 2s and 3s yet the overall rating was given as 2.

Conclusion

Review of the OPR documents stored on PRISM shows that:

- Information about the conduct of the review is generally not well presented. In particular, it is not easy to find which reviews are independent of DFID and which are done by staff. The involvement of stakeholders is poorly documented and review methodology is minimal.

- The nature of the OPR documentation and style of completion is extremely diverse. This applies not just to comparisons between full OPR documentation and the summary proforma, but to the content of these documents.

- Review practice varies widely. Some staff rate output delivery overall, others rate each output separately and others rate each output indicator. Assessment of the quality of purpose ratings is difficult because many purpose statements are poorly worded and are closer to outputs, and many of the purpose indicators lack a good specification. The quality of ratings based on results-chain categories such as project purpose is highly dependent on those concepts and indicators being well determined.

- The approach taken in drafting the reviews appears to be thorough and objective in a majority of cases. Clearly, the requirement to state the output and purpose indicators and report progress against them acts as a good stimulus for reporting.

ANNEX 3: KEY INFORMANT INTERVIEWS WITH DFID COUNTRY OFFICES

In order to better understand the processes that underpin project scoring, a number of telephone interviews were made to DFID Country Offices. Using a list of the most recent reviews (undertaken in the past two months), a list of the key Country Offices was compiled with assistance from DFID's Evaluation Department. The list consisted of: Bangladesh, Brazil, China, India, Malawi, South Africa, Tanzania, Uganda and Zambia. All Country Offices were contacted by email and followed up by telephone. In total, there were eight responses covering Bangladesh, Brazil, India and Malawi, including Advisors that cover Health, Engineering, Economics, Social Development and Rural Livelihoods. Additional meetings were held with DFID staff from the Evaluation, Performance Effectiveness and Statistics Departments.

Admittedly the sample size is relatively small, but given the constraints of this study, the primary purpose was to add an extra dimension to the evidence—and in particular to give insights into the way in which the review process (OPRs) is conducted in different Country Offices. The matrix below gives only tentative conclusions, which may need to be followed up in a more comprehensive study.

Key Narrative	Supporting Evidence
Overall: Country Offices have a need to measure performance for their own management purposes, and in general take a thorough approach to reviews. However:	All Country Offices and sectors interviewed explained comprehensive processes for reviewing projects, programmes and SWAps.
• reviews vary considerably, not necessarily between offices, but by sector, instrument and type of project.	While most offices said they undertook team reviews, the composition of the team and the review process varied immensely.
• innovations are occurring (often in isolation) with several Country Offices developing new ways to undertake reviews.	For example, DFID Malawi (Health) have recently completed a combined OPR of 3 related projects, and see this as a way forward for sector-wide approaches. Also, DFID Bangladesh are seeking to contract out all OPRs under one contract. With only one consultancy firm undertaking all OPRs, this is seen as a useful mechanism for enabling better consistency and understanding of DFID's approach.
• measuring value for money is very difficult, especially with the move towards sector-wide and budgetary support approaches.	In budgetary support, DFID Malawi (Economics) are developing a 'poverty-linked expenditure' approach for their budget support programme, where money is linked to key poverty reduction areas. Also, DFID Brazil have developed a matrix of outcomes and indicators showing how they intend to contribute to the VFM in the PSA.

Key Narrative	Supporting Evidence
Review Process: It is difficult to generalise the review process, as they tend to be tailored towards the particular needs of the projects and Country Offices. However:	All Country Offices described differing approaches to reviews, often depending on the size of the project, its nature (whether its problematic), the type of aid instruments, and the stage of implementation (annual review, mid-term, project completion).
• in general, reviews are undertaken by teams in conjunction with the partner (or implementing) agency.	All interviewees explained processes which were undertaken by teams in cooperation with partner agencies. This was described as a general pattern for most reviews.
• in some areas, there has been a move towards getting partners to 'own' and conduct the review process.	This is particularly important for SWAps where there may be several donor partners involved in a review. Other examples cited include getting country governments to be responsible for conducting the review, as part of institutionalising the project.
• there are many examples of OPRs being conducted with external consultants, in close cooperation with DFID advisors. This does however vary, with some sectors preferring not to use external consultants.	In addition to external (or independent) consultants, many Country Offices call upon the services of DFID Advisors in specialist areas such as Rural Livelihoods, Social Development, Economics, etc.
• in general, stakeholder consultation is seen as an important part of the review process, with field visits and beneficiary consultations also taking place for particular projects.	DFID India for example often use field visits as part of the Annual Review process, and this will typically involve speaking to beneficiary groups—though this depends on the exact nature of the project.
• reviews can sometimes be undertaken as separate mechanisms to the internal requirements of DFID—with the final report (the review) being used to complete OPR proformas, project scoring, etc. after the event.	For example, Country Offices may set out to have the review 'owned' by the implementing partners or at least jointly shared with the government staff. DFID staff then use the findings to complete internal procedures. On other occasions, external consultants will be former DFID Advisors who are able to complete internal procedures.
• the language and approach of OPRs is sometimes seen as being at odds with new ways of working (i.e. with the general trend to downsize the importance of projects).	The 'project-ised' approach of OPRs does not fit comfortably with sector programmes, SWAps and Budget Support. For example, the language of 'success/failure' does not capture a messy reality where performance may be in terms of a broader development process.
Project Scoring: project scoring is mostly seen as a central need, being of limited use for Country Office management purposes.	Some offices have used scoring for the occasional study to compare projects across sectors or across Country Offices. Others see scoring as a useful tool for focusing on the issues (especially the output ratings).

Annex 3

Key Narrative	Supporting Evidence
• There is a wide variety of ways of measuring performance, with Country Offices tending to measure in terms of indicators rather than overall scoring.	For example in DFID Malawi, SWAps may be measured using a set of process and impact indicators developed jointly with government and other partners. Also in Malawi, Budget Support is being increasingly linked to key input indicators (with limited outcome indicators), as the first priority is to ensure money is being spent in the key poverty areas.
• The quality of the review and scoring is usually only checked by the relevant Project Officer, and rarely by the Head of the Country Office.	Sometimes the Head of the Country Office will get involved when there are particular issues, or if they have a personal or professional interest in the project.
• Institutional pressure is difficult to assess, but there does not appear to be a particular pressure not to score projects 4 or 5.	While it is rare for a project to get a score 5, this may be partially because action is generally taken earlier in a failing project. In other circumstances projects can be extended or the log frame outputs changed.
Risks: Risk analysis is usually only done at the project design stage, and few Country Offices re-appraise the risk analysis once the project is up and running.	In general the risk analysis is seen as a one-off event. For some DFID offices it makes no sense to change the risk rating, while for others the possibility had not really been considered but might be worth pursuing.
• Risk may be assessed by the DFID Project Officer, sometimes jointly with an external consultant, while for others it is a wholly DFID staffed exercise.	The composition of the teams and methodologies for assessing risk vary considerably. Also, with limited guidance/tools, risk assessments can be rather unstructured.
• The risk categorisation and analysis (matrix) is almost always checked (and sometimes questioned) by the Country Office Head.	This seems to be the case across all Country Offices interviewed—and is largely because the Project Memorandum must go via the Head of the Country Office.

ANNEX 4: REFERENCES AND PEOPLE INTERVIEWED

DAC (2003), United Kingdom: Development Co-operation Review Main Findings and Recommendations , DAC Journal, Volume 2, No. 4.

DFID (2002a), Statistics on International Development 1997/98–2001/2, Department for International Development, UK.

DFID (2002b), Annex 1: Progress against targets in DFID's Public Service Agreement 2001/2–2003/4, extract from Autumn Performance Report, Department for International Development, UK.

DFID (2002c), Public Service Agreement, 2003–2006; and DFID Service Delivery Agreement, 2003–2006, Department for International Development, UK.

DFID (2002d), How Effective is DFID? Development Effectiveness Report (draft), Department for International Development, UK.

NAO (2002), Department for International Development. Performance Management—Helping to Reduce World Poverty, National Audit Office, UK.

MORI (2002), Public Service Reform: Measuring and Understanding Customer Satisfaction, A MORI review for the Office of Public Services Reform, MORI Social Research Institute.

List of people interviewed

- Joanne Asquith, Evaluation Department
- Matthew Sudders, SRSG, Performance Effectiveness Department
- Steve Martin, PRISM team, Performance Effectiveness Department
- Paul Whittingham, PDG, Performance Effectiveness Department
- Paul Marker, PDG, Performance Effectiveness Department
- George Holroyd, PDG, Performance Effectiveness Department
- Emily George, Statistics Department
- Liz Rolfe, Statistics Department
- Paul Ackroyd, Head of DFID Bangladesh
- Anne Austin, Health Advisor, DFID Malawi
- Peregrine Swann, Engineering Advisor, DFID Bangladesh
- Karl Livingstone, Economics Advisor, DFID Malawi
- Martin Leach, RNR Advisor, DFID Bangladesh
- Peter Evans, Social Development Advisor, DFID Malawi
- Hugh McGarvey, PRISM team, Performance Effectiveness Department
- Colin Kirk, Head, Evaluation Department
- Gail Marzetti, DFID Brazil
- Shiromani Singh, DFID India
- Asghar Ali, Evaluation Department
- Arthur Fagan, Evaluation Department

ANNEX 5: TERMS OF REFERENCE/SCOPE OF WORK

Purpose of the Study

The focus of this study is the PSA Value for Money indicator and the extent to which it is a relevant, efficient and effective measure of DFID's performance.

Background Information

As part of its Public Service Agreement (PSA), DFID is required to include an objective for increasing overall value for money (VFM) and an indicator for measuring progress against it.

The PSA states that value for money (VFM) is achieved in DFID when:

1. The proportion of DFID's bilateral programme going to low income countries increased from 78% to 90% and

2. There is a sustained increase in the index of DFID's bilateral projects' evaluated success. [41]

The first part of the indicator is relatively easy to measure. Its relevance to VFM is the assumption that more people can be lifted out of poverty per aid pound spent in the LICs than in middle and higher income countries. Hence increased VFM can be achieved through the reallocation of aid resources to the LICs.

While this is relevant for measuring aid inputs, it tells us little about the outputs and outcome of spending i.e. whether it is efficient and effective at achieving development objectives in the country where it is spent. It also assumes that all LICs are equally capable and equally efficient at reducing poverty.

To some degree, the second indicator helps to overcome this by measuring the extent to which projects and programmes are likely to achieve their development objectives (outcomes) and the risk involved in doing so. However, it is much more problematic in that it relates to the quality of expenditure rather than its quantity. It makes an explicit link between performance of the bilateral programme at the project and programme level (across sectors, countries and aid instruments) and overall value for money at the corporate level. The measure of VFM at the corporate level is hence dependent on the system for assessing the quality of projects and programmes financed from bilateral aid resources. This raises a number of questions and issues bout the relevance, efficiency and effectiveness of the indicator as an appropriate measure of value for money.

- **Relevance** i.e. the extent to which the indicator captures and covers DFID's main line of business (bilateral and multilateral) and emerging new ways of working.

- **Efficiency** i.e. the extent to which the indicator provides accurate and reliable information on performance. The degree of effort involved in collecting performance information, and the consistency of reporting across subject, time and space;

[41] See DFID PSA 2003-2006

- ➢ **Effectiveness** i.e. how useful is the indicator as a performance measure. What would a change in the indicator tell us about performance and how to further improve VFM across the organisation?

Study Approach and Methodology

Questions concerning relevance, efficiency and effectiveness will be answered by examining the quality of the data underpinning the value for money indicator.

The approach to linking project and programme performance and overall value for money has been criticised by external observers. The main issues concern:

- **Objectivity**: The extent to which the target itself could influence staff to score projects more highly than they otherwise might, undermining the quality and objectivity of project scoring. [42]

- **Verifiable**: Whether or not annual project scoring is consistently supported by a systematic project review process, which is rigorous and independently verifiable;

- **Consistent**: Whether the indicator itself is an accurate reflection of DFID value for money particularly when **aggregated** across all aid instruments, over sectors, countries and time; and

- **Reliable**: Whether risk is consistently and accurately assessed.

Methodological Issues

Performance data contained in PRISM will be analysed to:

- ➢ Identify patterns and emerging trends in performance over time, across regions, sectors and instruments; and
- ➢ Explain why and what is causing these to emerge.

This would require a more detailed look at the performance information captured in PRISM and the systems that underpin it. Project score and risk assessment data in particular would be analysed to assess:

- The proportion of scores supported by documentary evidence, or a record of how scores and risk were allocated;
- The proportion of scores supported by an OPR and those not;
- The proportion of OPRs conducted externally (i.e. by individuals not involved in the approval, management and implementation of project interventions); and
- The proportion of OPRs involving stakeholder participation (if this information is available)

Project scores that are not underpinned by OPRs would also be examined. Where possible telephone interviews would be conducted with DFID staff to assess the basis on which project scores

[42] Development Effectiveness Report (DFID Evaluation Department 2001). NAO: Performance Management: Helping to Reduce World Poverty" 2001

and risk ratings were made. The analysis would be repeated for risk ratings to see whether they are underpinned by rigorous risk assessments consistently applied across the organisation.

Both sets of data could be broken down by region. This would enable the identification of patterns and significant differences in project scoring and risk ratings across regions and the extent to which this is attributable to systemic weaknesses in performance assessment.

The second stage of the investigation would be based on a more detailed investigation of the systems and processes underpinning performance review, through detailed desk studies of the project and programme portfolio across regions, and by country visits. Recommendations to update and revise PCM guidance could then be made on the basis of these findings.

Issues for more detailed investigation include:

- The extent to which high project scores are supported by the achievement of log-frame outputs and outcomes as evidenced by the achievement of Objectively Verifiable Indicators (OVIs).
- The extent to which indicators have been clearly and accurately identified in log frames. This raises issues around the quality of project design and monitoring, and whether a more detailed quality assurance audit is needed to ensure quality concerns are systematically investigated.[43]

[43] e.g. Quality at Entry and Quality at Exit undertaken by the Quality Assurance Group in the World Bank

TERMS OF REFERENCE

Review of the VFM Indicator in DFID's Public Service Agreement

Purpose of the Review

The purpose of this study is to review DFID's Value for Money Indicator. The main focus will be on the performance data that is used to construct the indicator and the extent to which this is supported by objectively verifiable performance data.

Scope of Work

The issues to be addressed by the consultants are covered in the Approach paper appended to the terms of reference.

The consultants will be required to:

- Complete the first phase of the task as outlined in the approach paper methodology;
- Analyse project scoring and risk assessment data contained in PRISM to:
  (i) Identify emerging patters and trends; and
  (ii) Explain why these are occurring
- Review background information including the Technical note on the compilation of the VFM indicator, to look at past trends in the movement of the indicator and what this reveals about DFID performance
- Interview and consult DFID staffs that are engaged in monitoring DFID corporate performance and reporting this against the PSA (a list of relevant staff will be provided by EvD).

Output

The consultants will compile the evidence contained in PRISM in a way that is easily understood by the reader. Diagrams and tables should be used to identify and illustrate findings and trends. Colour should be used in charts and tables where this helps in identifying key trends.

A short paper (12 pages maximum) will be written on the reliability and objectivity of the performance information contained in PRISM. Progress in fulfilling the work outlined in the approach paper will be reported to DFID's Evaluation Department on 13th February in Abercrombie House. The final report would be completed by the end of February 2003.

A brief description on how quality assurance is managed by other donors could be useful in making recommendations on how the system could be strengthened.

Inputs

Up to 20 days of consultancy input is required. One consultant would need to be skilled in sorting data, analysing it and presenting it in a way that is easily understood.

Annex 6: Review of the VFM Indicator: Management Response

1. This study was prepared by independent consultants commissioned by DFID's Evaluation Department. The report was produced in April 2003; following internal consultation it was then submitted to me in October 2003. This management response was produced in November 2003.

2. DFID's mission is to contribute to the achievement of the United Nations Millennium Development Goals. These goals involve, inter alia, the achievement of a reduction of one-half in the proportion of people globally living in extreme poverty by 2015. DFID has a Public Service Agreement, like other UK Government Departments, setting out our corporate objectives and targets for the years 2003-2006. Our Public Service Agreement is designed around the contribution that we aim to make towards the achievement of the Millennium Development Goals.

3. Part of DFID's contribution to the achievement of these goals is delivered through the financing of poverty reduction projects and programmes through our country offices in developing countries. The causal links between the impact of activities we help finance and the achievement of the Millennium Development Goals are complex, and there are also complicated issues of aggregation. Nevertheless, we believe that the more effective our projects and programmes are in achieving their outputs and purpose, the greater will be DFID's contribution towards achieving our Public Service Agreement target and hence the Millennium Development Goals.

4. DFID has a range of systems for measuring the effectiveness of projects and programmes we finance. They include:
   a. A programme of evaluation studies, independently commissioned and managed by our Evaluation Department;
   b. A system of Project Completion Reports, under which, towards the end of the period of financing for any bilateral country activity to which we have made a commitment of at least £1m, we make a formal assessment of the extent to which the activity has achieved its intended objectives and what lessons there are to be learned;
   c. An annual system of scoring projects under implementation to which we are committing more than £1m, against the extent to which we expect them to achieve their intended purpose and output.

5. DFID's Public Service Agreement includes a value-for-money indicator related to the impact of our spending activities as reflected in the data collected under the third of these systems (project scoring during implementation). We have a target to achieve a sustained increase in the value-for-money indicator, which is disaggregated by risk category. We report on progress against this indicator annually in our Departmental Report which is presented to Parliament (and we collect information on it quarterly in a management report submitted to DFID's Management Board).

6. The scoring of projects and programmes we finance which are covered by the value-for-money indicator is undertaken by DFID staff. In many cases, the information on which they base their judgements include monitoring and other reports prepared by outside independent specialists

and/or with the collaboration of the Government of the country in which the activity is taking place. Nevertheless, because the scoring system involves judgements made by our own staff, questions potentially arise as to the scope for errors or bias in the scoring. One of the objectives of this study was to assess whether there was any evidence of any systematic error or bias in the scoring.

7. The main conclusions DFID senior management draw from the study, and the further action we plan, are:

   a. that the system of project-scoring during implementation has bedded down in a reasonably satisfactory way over a short period. There appears to be no reason to believe on the basis of the evidence currently available that there is any systematic bias in scoring;
   b. internal compliance with the scoring system has increased substantially since 2001. As at the third quarter of 2003, 92% of eligible bilateral country projects have been scored within the last twelve months. This means that the database of scored projects is substantial, at over 440 bilateral activities currently;
   c. the proportion of DFID's bilateral country programme portfolio by value that is scored is also substantial, at 70% by commitment value. Nevertheless, because DFID finances a large number of projects to which our commitment level is less than £1m, the proportion of the portfolio scored by number of project is lower, at around 15%. DFID is content that this level of coverage is currently acceptable, but we will keep under review whether we need to look in future in more detail at a sample of projects of lower value;
   d. we have noted the variations in risk categorisation across region of the value-for-money indices. We agree that this raises issues for management to consider. As a first step, we have instigated in our 2003 review of Divisional Delivery Plans, a system of moderating risks across regions, and we have also provided additional guidance to spending departments on risk assessment at the project level;
   e. while the report concludes that supporting evidence and the quality of reviews underlining the scoring is reasonably thorough, we agree that it would be desirable to strengthen the evidence base further. We will therefore undertake a programme of more detailed independent reviews of selected projects and programmes with a view to providing further evidence on the quality of performance management and the impact of DFID spending activities at the project level;
   f. we also agree with the observation in the report that the scoring system should be supplemented by other means of assessing the effectiveness of very large allocations of financial aid, for example budgetary support. DFID's Evaluation Department is carrying out, in collaboration with a number of other major development agencies, a substantial programme of work to evaluate the impact of budget support operations. This will buttress our knowledge of the effectiveness of such expenditure.

8. We will also consider further the need for follow-up studies to the current one in the light of evolving evidence on trends in the value-for-money indices.

Mark Lowcock
Director General, Corporate Performance & Knowledge Sharing

5 November 2003

Annex 7: Data on scores for achievement and risk as at November 2003

a) Scores Distribution Over 12 Months from 1/11/02 – 1/11/03: Africa and Asia

Africa						
Rating	1	2	3	4	5	X
No/ projs						
(185)	14	71	49	14	0	38
%age	**8**	**38**	**26**	**8**	**0**	**20**
Asia						
Rating	1	2	3	4	5	X
No/projs						
(168)	10	69	46	15	3	25
%	**6**	**41**	**27**	**9**	**2**	**15**

b) Risk Level Distribution as at November 2003

		High Risk			
		Africa	**Asia**	**EMAD**	**Others**
No. of Projects:	**125**	40	41	27	17
% of Total		32	32.8	21.6	13.6
Commitment Value £m	**860**	357	341	57	105
% of Total		41.51	39.65	6.63	12.21
		Medium risk			
No. of Projects:	**768**	279	185	179	125
% of Total		36.33	24.09	23.31	16.28
Commitment Value £m	**4976**	1520	1728	291	1437
% of Total		30.55	34.73	5.85	28.88
		Low Risk			
No. of Projects:	**922**	228	155	150	389
Commitment Value £m	**1524**	386	516	110	512
% of Total		25.33	33.86	7.22	33.60

THE DEPARTMENT FOR INTERNATIONAL DEVELOPMENT

The Department for International Development (DFID) is the UK Government department responsible for promoting sustainable development and reducing poverty. The central focus of the Government's policy, based on the 1997 and 2000 White Papers on International Development, is a commitment to the internationally agreed Millennium Development Goals, to be achieved by 2015. These seek to:

- Eradicate extreme poverty and hunger
- Achieve universal primary education
- Promote gender equality and empower women
- Reduce child mortality
- Improve maternal health
- Combat HIV/AIDS, malaria and other diseases
- Ensure environmental sustainability
- Develop a global partnership for development

DFID's assistance is concentrated in the poorest countries of sub-Saharan Africa and Asia, but also contributes to poverty reduction and sustainable development in middle-income countries, including those in Latin America and Eastern Europe.

DFID works in partnership with governments committed to the Millennium Development Goals, with civil society, the private sector and the research community. It also works with multilateral institutions, including the World Bank, United Nations agencies, and the European Commission.

DFID has headquarters in London and East Kilbride, offices in many developing countries, and staff based in British embassies and high commissions around the world. DFID's headquarters are located at:

DFID
1 Palace Street
London SW1E 5HE
UK

and at:

DFID
Abercrombie House
Eaglesham Road
East Kilbride
Glasgow G75 8EA
UK

Switchboard: 020 7023 0000 Fax: 020 7023 0016
Website: www.dfid.gov.uk
email: enquiry@dfid.gov.uk
Public enquiry point: 0845 3004100
From overseas: +44 1355 84 3132

ISBN 1 86192 597 2

Produced by AKROS PRINT LTD for DFID International Division