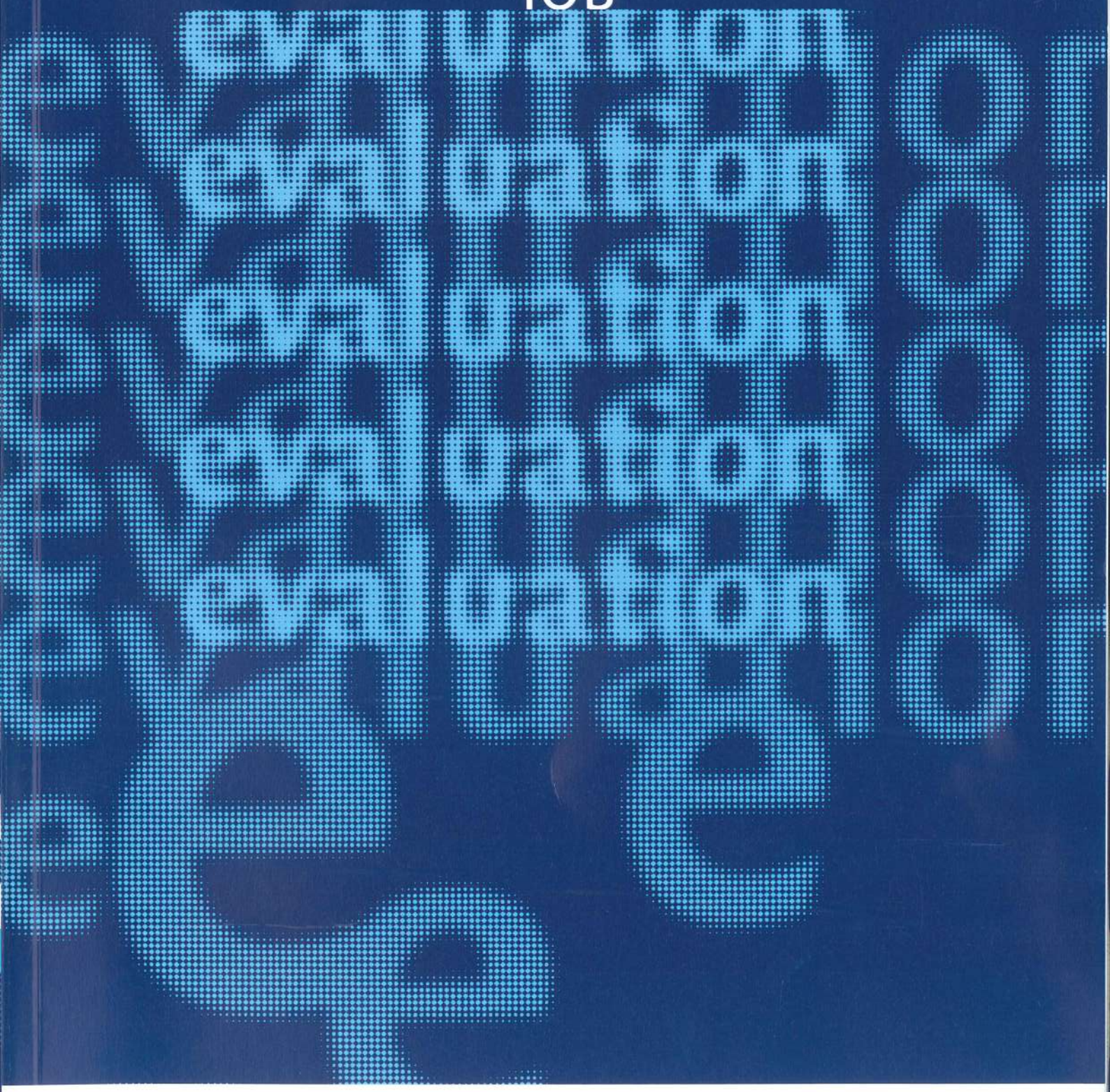
IOB



QUALITY ASSESSMENT OF EVALUATIONS COMPLETED IN 2002

FINAL REPORT

IOB EVALUATIONS No 298 | AUGUST 2004

POLICY AND OPERATIONS EVALUATION DEPARTMENT

IOB EVALUATIONS | No 298

Policy and Operations Evaluation Department | Ministry of Foreign Affairs, The Netherlands

Quality Assessment of Evaluations completed in 2002

Final Report

IOB | August 2004

PREFACE

Dozens of evaluations are carried out within the Netherlands Ministry of Foreign Affairs every year. The purpose of these evaluations is to probe the effectiveness and efficiency of the policy the Ministry has pursued and, specifically, to answer the questions: Did policy have the intended effect and was it properly implemented? Evaluations have a direct association with the government's result-oriented management model that forms the basis of the 'From Policy Budgets to Policy Accountability' initiative.

Apart from financial information the Ministry's budget and annual report increasingly contain information on what the Ministry does and has done, the effects it intended to achieve and what it actually did achieve. This information contributes to sound decision-making on the formulation and implementation of policy. It is therefore vital that the evaluations from which this information is obtained are based on valid and reliable investigations that produce useful results.

The Ministry of Foreign Affairs' central policy evaluations are carried out by the Policy and Operations Evaluation Department (IOB). Evaluations are also carried out decentrally by policy departments and missions, in regard to the policy and programmes for which they are responsible. IOB has now carried out a quality assessment of the decentral evaluations completed in 2002 and this report presents the results of that assessment. The assessment was based on the quality requirements set out in the Order on Performance Data and Evaluations in Central Government.

This assessment is the first in a series of annual assessments that will examine the quality of a selection of evaluations carried out during the preceding year. The present assessment sought chiefly to assess the quality of evaluations and to identify omissions and possible improvements. The assessment is important not only because it supports the process of improving the quality of decentral evaluations but also because it contributes to the knowledge of the reliability of policy information used by the Ministry of Foreign Affairs.

The assessment was carried out by Rita Tesselaar, IOB Inspector, and Roland P.A. Rodts, consultant. The IOB would like to express its sincere thanks to the many persons, too numerous to mention, whose knowledge, experience and comments have made an indispensable contribution to the assessment. Their contribution does not mean that they bear responsibility for the report's contents; that responsibility rests entirely with this Department.

Rob D. van den Berg
Director, Policy and Operations Evaluation Department

Abbreviations

BZ	Netherlands Ministry of Foreign Affairs
DGIS	Directorate General International Co-operation
RPE	Order on Performance Data and Evaluations in Central Government
VBTB	From Policy Budgets to Policy Accountability
FEZ	Financial and Economic Affairs Department
IOB	Policy and Operations Evaluation Department
HBBZ	Operational Procedures Manual

CONTENTS

1. MAIN FINDINGS AND CONCERNS

1.1 Introduction

The Order on Performance Data and Evaluations in Central Government (the "RPE") has been in force since 2002. It sets a number of quality requirements for the instruments used to evaluate the policies that underlie ministerial budgets and annual reports, which must be valid, reliable and useful. It also stipulates that the quality of those evaluations must be subject to sample based verification. The Policy and Operations Evaluation Department (IOB) of the Netherlands Ministry of Foreign Affairs (BZ) is charged with the assessment of the quality of evaluations carried out by the Ministry's departments and missions. The present assessment relates to evaluations completed in 2002. Its central objectives are to assess the quality of the evaluations and to identify omissions and possible improvements.

The RPE implements the provision in the Government Accounts Act that states that Ministers are responsible for the effectiveness and efficiency of the policy on which their budgets are based. According to the Act, Ministers must make periodic checks. A specific objective of the RPE is to "ensure that the policy information provided in the context of the ministerial budget and the annual report satisfies the quality requirements applicable to such information". The RPE grew out of the aim, contained in the government policy document "From Policy Budgets to Policy Accountability" (VBTB), to foster result-oriented management. This assessment of the quality of evaluations is part of a longer-term process of putting the VBTB and RPE into effect.

The assessment relates to a sample of 25 of the total of 83 evaluations completed in 2002. The majority of these evaluations were of programmes and projects in the international co-operation field. An attempt was made to objectify the quality assessment as far as possible by using a list of 16 indicators and associated components. The indicators are divided into three clusters, reflecting the quality requirements set out in the RPE (validity, reliability and usefulness) and are assessed on a four-point scale (Very Good, Good, Fair and Poor). Two assessment approaches were used, a desk study of all 25 evaluations and an additional field study of 10 of them.

1.2 Main findings

- *Decentral evaluations provide only a limited contribution to information on the effectiveness and efficiency of policy and the efficiency of operational management*

In the evaluations that were examined the question as to effectiveness relates largely to whether operational objectives at the level of the implementing organisation were achieved (the organisation's performance or "output"), and much less to the *effects* of that output. Around three-quarters of the evaluations fail to address the issue of efficiency or devote very little attention to it. Evaluations are largely designed from a managerial and future-oriented perspective and in most cases an assumption is made that the programme or project in question will continue. The evaluations are designed much less from a policy perspective, i.e. with a view to identifying the effects of policy. Nor do they form part of a broader

ministerial effort to systematically assess effectiveness and efficiency and use the information so obtained for budget and annual report purposes.

- *The evaluations are of variable quality*

The table below shows how many of the 16 indicators received the score Very Good or Good for the 25 evaluations examined.

Table showing "Very Good" and "Good" ratings

Number of "Good" or "Very Good" ratings	13-15	9-12	5-8	3-4
Number of evaluations	6	8	9	2

It can be seen from the table that 6 evaluations were rated "Good" or "Very Good" on all points. There is a middle group of 8 evaluations in which these higher ratings were awarded to between 9 and 12 indicators and there are a further 11 evaluations that may be regarded as Fair or Poor in quality.

Indicators on which a large number of evaluations received a Good or Very Good rating are: the description of the reasons for and purpose of the evaluation, consistency between the questions, findings, conclusions and recommendations, trouble-free implementation of the investigative work and the match between the intended use of the evaluation and the evaluation results.

The following common shortcomings were identified in regard to the three quality requirements of validity, reliability and usefulness.

I. Validity

The main shortcomings in regard to validity were the failure to place the subject of the evaluation clearly enough in its policy and institutional context and deficiencies in the description of the subject of the evaluation. Information is often incomplete or even absent, particularly information about programme/project inputs and the way that funding was spent. The soundness of the analysis is frequently impaired by a failure to specify outputs and/or effect indicators with sufficient clarity. The evaluations also devote little attention to measuring results. In many cases the results achieved are not set out systematically and/or no clear distinction is made between different kinds of results (actual inputs, performance, effects). The quality of the efficiency analyses also leaves much to be desired.

II. Reliability

The reliability of many evaluations is impaired because, of all those who have an interest in the subject under evaluation, only programme or project managers and implementers were interviewed, or because the sample taken was too small. Nor is much attention devoted to the quality of the information on which the evaluation is based. The independence of evaluation teams is under pressure because evaluators have been selected and paid for by the budget holder or someone else who has a direct interest. Furthermore, the quality of the evaluations has not been subjected to systematic checks by the department or mission concerned.

III. Usefulness

The RPE states that the usefulness of an evaluation depends to a great extent on its validity and reliability. Deficiencies in validity and reliability will, of course, impair an evaluation's usefulness. In addition, a large proportion of the evaluations do not

answer all of the evaluation questions set out in the Terms of Reference. The way the evaluation results are presented also leaves a lot to be desired in many cases.

The quality assessment did not look in any depth at the use made of the evaluations because this is not in itself a good indicator of quality. The assessment would need to be designed differently to gather the necessary information. The use made of evaluations was, however, a point of concern in the field studies. In most cases the activities evaluated were to be continued and the interviewees stated that the evaluation results had played a greater or lesser part in the decision on whether to continue the activity, but it was not possible to establish this in detail and with certainty.

- The failings identified are due to methodological, policy and process-related factors

The evaluations are not based directly on the RPE and have specific purposes. Because they concentrate to a lesser degree than the RPE on analysing effectiveness and efficiency, applying an assessment framework based on the RPE results in low ratings for some of the indicators.

Changing policy priorities and pressure on spending contribute to a concentration on operational aspects and future orientation in the evaluations. The Terms of Reference of 23 of the evaluations contain future-oriented questions from which it can be deduced that there was an intention, when the Terms of Reference were drawn up, to continue the activity. This future-oriented perspective means that the staff concerned regard evaluation as a different type of instrument from the ex-post evaluation instrument defined in the RPE.

The staff concerned (policy officers, managers, controllers) are not completely familiar with the principles and frameworks on which evaluation is based. Some of them are not familiar with the Evaluation Guide from the Operational Procedures Manual. Only one or two are familiar with the RPE. Apart from this, factors such as vagueness of the Terms of Reference, pressure of time when selecting consultants, the lack of predetermined output/effect indicators and of monitoring data, and overburdening of the work programme also contribute to the failings.

No connection was established between the professionalism of evaluators and the quality of evaluations. Professionalism was examined on the basis of the knowledge and experience set out in the curricula vitae. All the curricula vitae scored Very Good or Good on professionalism.

On average, evaluations organised in the multi-donor context score Very Good or Good more often than bilateral evaluations. The multi-donor evaluations were characterised by more precisely formulated Terms of Reference, a relatively large evaluation budget, careful selection of consultants with international experience, intensive monitoring and support of the evaluation process, and wide dissemination of the evaluation results.

1.3 Concerns

Based on the findings a number of problems were identified, the most important of which are:

I. Different types of investigation

The assessment shows that departments and missions do not make any systematic distinction between investigations focusing on operational aspects – which are referred to in the Evaluation Guide as reviews[1] – and evaluations, which look principally at effects, effectiveness and efficiency.

II. Different types and levels of results

Many evaluation reports do not make a clear distinction between different types and levels of results, the planned and actual input/expenditure, direct results (performance or output) and achievement of the objectives (effects or outcome). Information on the different levels of results is necessary to be able to make judgements about effectiveness and efficiency.

III. Extent to which activities can be evaluated

Evaluation at the various result levels is made more difficult by the absence of measurable objectives and because the intended results and result indicators have not been clearly defined, or defined at all, when activities were planned.

IV. Evaluation in the context of the policy cycle

Only to a very limited extent do evaluations form part of a broader effort by policy departments to systematically examine, assess and promote effectiveness and efficiency as part of the policy cycle.

V. Minimum framework for the quality of evaluations

The fact that the Evaluation Guide contained in the HBBZ is not binding on evaluators is a factor in inconsistent application of evaluation principles and frameworks. Detailed regulation is not an appropriate response to this problem because evaluations largely demand a tailored approach. What is lacking, however, is a minimum framework for evaluations, setting out the evaluation principles and laying down minimum requirements for Terms of Reference, selection of consultants, the content and completeness of reports.

VI. Knowledge of evaluation

The findings show inadequate knowledge about evaluation and inadequacies in the application of such knowledge. Various measures can be conceived to solve this problem, such as ways of transferring know-how, a minimum framework for the quality of evaluations in the shape of an instruction or concise handbook, a guide for departments and missions for quality checks on evaluations, and expert support. However, none of these solutions is capable of eliminating shortcomings in the quality of evaluations, unless management at the various levels motivates the staff concerned to devote the necessary energy to them.

VII. Assessment of quality

The set of indicators (based on the RPE quality requirements) that was used for this assessment was found to be useful. Indicators that relate to the evaluation criterion of efficiency, to the representativeness of the evaluation, verification of the

[1] The Evaluation Guide defines reviews as instruments that focus on operational aspects of programme or project implementation (programme/project design, implementation structure, progress of implementation, etc.).

information sources and the independence of the evaluation team could be further refined if the Ministry were to formulate minimum requirements for these matters.

Legislation and regulations require the authorities to provide information on the *efficiency* of both policy (by examining the relationship between cost and the effects achieved) and operations (by examining the cost and quality of the products/services or output provided). It is clear from the assessment that the use and analysis of this criterion leave much to be desired. Although many of the activities do not lend themselves to a quantitative approach (cost-benefit analysis), even a less comprehensive analysis is often lacking due to the absence of assessment criteria and standards.

The RPE states that "the units to be investigated must be selected and assembled in an appropriate way, e.g. drawn at random from a sufficiently large sample". It is not clear from this when the requirement for *representativeness* has been satisfied. Statistical methods for representative research are often unusable in BZ's policy fields because the information on the research population required by such methods is not available.

The RPE stipulates that the sources of information used must, wherever possible, be independent. In practice, evaluators largely restrict themselves to information provided by persons directly involved (e.g. annual reports, monitoring data, minutes of meetings, internal evaluations). In most evaluations *verification of this information* is a difficult issue. Although there is virtually always some form of triangulation, inasmuch as evaluators have spoken to a number of different concerned parties, it is no easy task to determine to what extent this results in a true picture.

Where evaluators are selected and paid for by a department or mission that has responsibility for the matter under evaluation, complete *independence of the evaluation teams* is unachievable. In such cases a degree of distortion of the evaluation results cannot be completely ruled out.

2. BACKGROUND TO THE ASSESSMENT

2.1 Reason for and purpose of the assessment

There are national quality standards for the tools used to evaluate the Ministry of Foreign Affairs' (BZ) budget and accounts. These standards are set out in the Order on Performance Data and Evaluations in Central Government (RPE), which has been in force since 2002.

The Policy and Operations Evaluation Department (IOB) has the task of assessing the quality of evaluations completed by BZ departments and missions, on a sample basis. Starting in 2003 and annually thereafter IOB will carry out quality assessments of a sample of the evaluations completed during the preceding year. The current assessment is the first in the series and relates to evaluations completed in 2002. Its main objectives are to assess the quality of the evaluations and to identify any omissions and areas for improvement.

2.2 Background to the assessment

Evaluations carried out at the Netherlands Ministry of Foreign Affairs are divided into:

I. evaluations by IOB, which focus primarily on specific policy themes, instruments or programmes. These are termed 'central evaluations';

II. evaluations by the Ministry's departments or missions. These evaluations have so far related mainly to specific activities (primarily projects and programmes) and, to a lesser extent, themes and sectors. These are termed 'decentralised evaluations'. The present assessment is concerned with the quality of decentralised evaluations.

A guide to evaluations is given in the Ministry's Procedural Guidelines. A lack of information on the quality of decentralised evaluations led IOB to carry out its first assessment of evaluation and monitoring in the bilateral development cooperation area in 1993.[2] The findings indicated inadequacies in the evaluations, which meant that evaluation and monitoring were less effective than they could be. The main findings were that no systematic use was made of evaluation results in formulating policy, methodological requirements were not satisfied, there was little involvement of partner organisations and local evaluators, and the role of monitoring was not clear.

At national level the government observed in the 1990s that insufficient attention was being paid to the way in which action by public authorities was evaluated. This led to the establishment of the Interministerial Supervisory Committee on Performance Data and Policy Evaluation (IBP). The IBP was charged with drafting a new Order to clarify evaluation standards. The resulting Order on Performance Data and Evaluations in Central Government (RPE) was developed during 2001 and came into force on 1 January 2002.

The aim of the RPE is 'to ensure that evaluations carried out within central government are sufficiently robust and that policy information provided in the context

[2] For the results of that evaluation, see: Evaluation and Monitoring, summary evaluation report 1995.

of ministry budgets and annual reports meets the quality standards relevant to this type of information'.

The RPE:

I. implements the provision in the Government Accounts Act which states that ministers are responsible for the effectiveness and efficiency of the policy on which their budgets are based and that they should make regular checks;

II. is in line with the government policy document 'From Policy Budgets to Policy Accountability' (VBTB) and is based specifically on that document's aim of giving impetus to the result-oriented management model;

III. distinguishes between performance data systems and evaluations and, within the latter category, between evaluations *ex ante*, *ex post* and operational assessments;

IV. contains further rules on the integrated use of evaluation tools, the extent to which and frequency with which policy is covered by regular *ex post* evaluation and on the technical and methodological quality of evaluation tools.

In 2001-2002, with a view to implementing the RPE, IOB and the Financial and Economic Affairs Department (FEZ) commissioned an assessment of the quality of evaluation reports relating to evaluations completed during the period 1997-2001. The main conclusion of this assessment was that "the evaluation reports do not adequately fulfil the quality standards set out in the Order on Performance Data and Evaluations in Central Government (validity, reliability and utility)". The most significant shortcomings of the evaluation reports were: insufficient information was provided about the evaluation, insufficient attention was given to the quality of the information on which the evaluations were based, evaluations lacked internal consistency, and insufficient attention was paid to the communicative aspects of the reports.

In April 2002 the memorandum 'BZ Policy on Organisation of the Evaluation Function: Putting the Order on Performance Data and Evaluations into effect' was adopted by the Ministry's senior management. The memorandum sets out the responsibilities for the BZ evaluation function: the director or director-general responsible for the relevant policy article determines the optimum and relevant level of coverage required from the evaluation and sets the timetable, while FEZ is responsible for coordinating and promoting evaluation work. Under the Government Accounts Act FEZ must also report annually to the Netherlands Court of Audit and the Finance Ministry on the investigations into policy effectiveness and efficiency. To this end, FEZ draws up BZ's budget-related evaluation programme each year. The recently established Audit Committee is responsible for central control of BZ's evaluations and, in that context, for making random checks to verify the periodicity, quality and use or potential use of evaluations. The memorandum confirms IOB's task in assessing the quality of evaluations.

2.3 Scope of the assessment

A list of decentralised evaluations completed in 2002 was drawn up, based on the evaluations reported by missions and departments in the evaluation annex to the 2003 annual plan. A total of 83 evaluations were reported – 27 by ministry departments and 56 by missions.

These evaluations are not evenly spread across articles, departments and missions. There are six policy articles for which no evaluations were reported. Evaluations by departments centre on article 6 (bilateral development cooperation), article 10 (co-operation with civil society organisations) and article 12 (co-operation with the private sector), all within the field of international cooperation.

Table 1 Evaluations completed in 2002 and reported by departments (broken down by policy article)

Dept.	art. 1	art 4	art. 6	art. 9	art. 10	art. 11	art. 12	art. 13	art. 14	art. 18	Total
DMV	1	1									2
DAO			1								1
DCO					5						5
DSI					9	2					11
DVF				1							1
DDE							4				4
CBI							1				1
DPV									1		1
FEZ										1	1
Total	1	1	1	1	14	2	5		1	1	27

Article 1 International order
Article 4 Good governance, human rights and peacebuilding
Article 6 Bilateral development cooperation
Article 9 International Financial Institutions
Article 10 Cooperation with civil society organisations
Article 11 International education
Article 12 Cooperation with the private sector
Article 13 Promotion of political and economic interests
Article 14 Asylum, migration and consular services
Article 18 General

A total of 22 missions reported one or more evaluations. As Table 2 below shows, the 56 evaluations reported are mainly in the field of bilateral development cooperation (article 6).

Table 2 Evaluations completed in 2002 and reported by missions (broken down by policy article)

Mission	Art. 1	Art. 4	Art. 6	Art. 9	Art. 10	Art. 11	Art. 12	Art. 13	Art. 14	Art. 18	Total
Bamako			4								4
Bogotá			5								5
Colombo			1								1
Cotonou			1								1
Dakar			1								2
Dar es Salaam			2								2
Dhaka			1								1
Guatemala			7								7
Kampala			2								1
Khartoum			1								1
Kigali			2								2
La Paz			2								2
Lusaka			3								3
Madrid								1			1
Managua			2								2
Maputo			6								6
Nairobi			3								3
Ouagadougou			4								4
Pretoria			1								1
Paramaribo			1								1
Sanaa			3								
San José		1	1								4
Sarajevo		1									2
Total		2	53					1			56

Article 1 International order
Article 4 Good governance, human rights and peacebuilding
Article 6 Bilateral development cooperation
Article 9 International Financial Institutions
Article 10 Cooperation with civil society organisations
Article 11 International education
Article 12 Cooperation with the private sector
Article 13 Promotion of political and economic interests
Article 14 Asylum, migration and consular services
Article 18 General

In order to make a precise selection of evaluations for the quality assessment, a number of basic data from all 83 completed evaluations were analysed, i.e. the relevant policy article, the department or mission responsible, the type of subject being evaluated (project, programme, sector, theme, other), the type of evaluation (final evaluation, interim evaluation, review, evaluation/formulation, other) and commissioning body (internal, joint, multi-donor). From the total of 83 evaluations, 25 were selected, taking account of: (i) the spread across policy articles, (ii) the spread across departments and missions and (iii) the spread across different types of evaluation. The evaluations selected comprise ten from ministry departments: DCO (4), DSI (2), DDE (2), DMV (1) and DAO (1), and 15 that were reported by missions: Bamako (1), Bogotá (1), Colombo (1), Cotonou (1), Dhaka (1), Kampala (1), La Paz (1), Pretoria (1), Sana'a (1), Maputo (3) and Nairobi (3).

2.4 The assessment questions

As stated in the Terms of Reference (Annex 1), the key questions the assessment seeks to answer are:

What happened in the area of decentralised evaluations in 2002? The assessment started by identifying the main characteristics of evaluations, such as the type of subject (theme, sector, programme, project, organisation, process, etc.); the subject to which they related (description, BZ's financial contribution, period covered); the type of evaluation (final evaluation, interim evaluation, evaluation/formulation, etc.); cost and BZ's contribution thereto; the commissioning body (budget holder, joint, multi-donor); and who carried out the evaluation (internal or external implementers).

I. To what extent do the evaluations completed in 2002 comply with RPE quality standards? The central question here is: to what extent are the evaluations valid, reliable and useful? A further point looked at was whether evaluations have a policy function – i.e. whether any link was made with policy objectives and whether evaluation results feed back into policy.

II. What factors affected the quality of the evaluations and how? The assessment examined which factors help explain the quality of the evaluations.

The answers to these questions were used to identify areas for improvement.

2.5 Assessment approach and methods

A twofold approach was adopted for the assessment: a desk study and a field study.

- *Desk study*

The desk study looked at the characteristics of the 25 selected evaluations and analysed and rated the evaluations, based on the Terms of Reference and the relevant evaluation reports. This phase consisted of:

I. testing the usability of the list of characteristics, criteria and indicators drawn up for the analysis. The list was revised and refined in response to the findings;

II. analysis and initial assessment of the selected 25 evaluations by studying the Terms of Reference and the evaluation reports themselves, based on the list of characteristics and indicators drawn up for that purpose;

III. making adjustments in response to comments made by the department or mission concerned on the list of characteristics, scores and supporting notes;

IV. identification of gaps in the data, for the purpose of the complementary field study;

V. on the basis of the above, drawing up a checklist of data to be collected for the field study (see Annex 3).

- *Field study*

The object of the field study was to verify the findings of the desk study and to fill in gaps in the information, particularly in regard to factors which affect the quality of evaluations.

A total of 10 of the 25 evaluations were selected for the field study, four of them being evaluations by policy departments (two each from DCO and DSI) and six evaluations by missions (three each from the Maputo and Nairobi missions).

The field study consisted of:

I. preparations for the field study, which entailed drawing up a list (per selected evaluation) of parties who would be involved in it, a list of interviews to be conducted, and the timetable;

II. execution of the field study, with a view to getting a detailed picture of the preparations for and execution of the evaluations in question and of the use made of feedback from them. Information was gathered by studying relevant files and by conducting interviews with the staff concerned, with the persons who commissioned the evaluations, implementers/managers, evaluators, members of evaluation steering committees or supervisory groups and other interested parties. The field study took place from 17 January to 6 March 2004. A list of persons interviewed is given in Annex 6;

III. reporting on the findings of the field study and taking account of the comments of those involved. The three reports in question (relating to Maputo, Nairobi and DSI/DCO) are included in the quality assessment dossier;

IV. adding to and/or adjusting the analysis and ratings of the evaluations on the basis of findings from the field study.

2.6 Limitations of the assessment

This is the first in the series of quality assessments of evaluations to be carried out annually on the basis of RPE quality standards. Indicators for these standards are still being developed at central government level. The list of indicators drawn up for the assessment was found to be fitted for its purpose but consideration should be given to possibilities for further improvement of the list of indicators.

Efforts were made to make the quality assessment objective by using a set of indicators and associated components that were defined as precisely as possible. However, in many instances it is not possible to determine objectively to what extent the indicators are satisfied. Consequently there may be differences in ratings due to the individual interpretations or perceptions of the assessors.

Although the assessment was not completely representative, it nevertheless covered a fairly broad spectrum of completed evaluations. Because of the diversity of the evaluations a representative assessment would have required a very large sample. This option was discounted, given the limited scope of many of the evaluations in comparison with the likely cost and added value of the quality assessment.

When the evaluations were examined the emphasis was on those aspects of quality that relate to the structure, implementation and reporting of evaluations. No pronouncement is made as to the accuracy of the information drawn from the evaluations, although the assessment did examine the depth and scope of such information.

The quality of the assessment was somewhat impaired by differences in the information available about the evaluations. Where evaluations were included in the field study (10 in total) it was possible for the assessors to fill in the information gaps identified in the desk study by their own investigations. In the remaining 15 evaluations the findings of the desk study were submitted to the mission or departments responsible with a request for them to comment. In six cases a written response was received that led to additions and/or adjustments being made to our

assessment of the evaluation. In the other nine cases the assessment is based solely on the desk analysis of the relevant Terms of Reference and the evaluation report.

The fact that the field study was limited to 10 evaluations means that the findings on factors that affect the quality of evaluations should be regarded as illustrative only.

2.7 Structure of the report

The final report contains three chapters. Chapter 1 sets out the main findings and areas for improvement. Chapter 2 explains the background to and objectives of the assessment, the approach adopted and the methods employed. Chapter 3 highlights the most significant characteristics of the selected evaluations and reports the findings of the quality assessment. These findings are arranged according to the three quality standards required by the RPE: validity, reliability and utility. It concludes with a final summary of the findings and an analysis of factors that affect quality.

3. QUALITY OF EVALUATIONS

3.1 General characteristics

It is clear from examination of the 83 evaluation reports completed in 2002 that the majority of the evaluations relate to individual activities (programmes or projects) for which the department or mission concerned is the budget holder. These are largely interim and final evaluations of projects or programmes and form part of the regular 2-4 year cycle that is a characteristic of activities financed by BZ. Around 25% of the evaluations were performed by third parties or in partnership with third parties.

As reported in Chapter 2, the quality assessment relates to a selection of 25 of these 83 evaluations. The table below shows their titles, dates and responsible budget holders, with the ten covered by the field study being indicated in **bold**.

The group of evaluations that relate to policy article 6 (bilateral development cooperation) accounts for 16 evaluations, representing around 65% of the sample. All but one of these evaluations were reported by missions. The remainder are evaluations by policy departments, spread across policy articles 1, 10, 11 and 12.

Table 3 List of evaluations selected

No.	Policy article	Title	Date	Dept./ Mission
1	Art. 1	Programma Ondersteuning Buitenlands Beleid voor mensenrechten	Jan 2002	DMV
2	Art. 6	Evaluation Asia Facility	Nov 2002	DAO
3		Programme de Développement Sanitaire et Social, Mission d' évaluation externe	Nov 2002	Bamako
4		Programa Estrategias para la Consolidación y Fortelicimiento del Sistema de Parques Nacionales Naturales, Primera Misión de Monitoreo y Evaluación	March 2003	Bogotá
5		Centre for the Study of Human Rights, evaluation of the 1997-2001 project	May 2002	Colombo
6		Evaluation externe de la mise en oeuvre des Programmes de l'Accord sur le Développement Durable par le CBDD au Bénin	Sept. 2002	Cotonou
7		Mid-term Review Char Development and Settlement Project II		Dhaka
8		Mid-term evaluation of the Technical Assistance provided under the Royal Netherlands Embassy sponsored District Local Government Support Programme	May 2002	Kampala
9		Evaluación del Fondo de Alivio a la Pobreza.	Sept. 2002	La Paz
10		**PROAGRI Evaluation**	**Dec. 2002**	**Maputo**
11		**Sistemas de Aguas (SAS) Evaluation**	**Feb. 2002**	**Maputo**
12		**Avaliação Intermedia da UDEBA na Segunda Fase 2000-2002**	**Sept. 2002**	**Maputo**
13		**Systems Evaluation of the National Civic Education Programme**	**Nov. 2002**	**Nairobi**
14		**External Evaluation of the Semi-arid rural development programme**	**April 2002**	**Nairobi**
15		**Mid-term Review of the Engendering the Political Process Programme**	**Nov. 2002**	**Nairobi**
16		RNE Support to the Youth Sector in South Africa	June 2002	Pretoria
17		Final Evaluation Yemen Drug Action Programme	Sept. 2002	Sana'a
18	Art. 10	Evaluation du Programme de Recherche Delta Niger-Mali	Aug. 2002	DCO
19		Evaluation of phase II and development orientations for phase III, Vietnam Netherlands Research Programme	Oct. 2002	DCO
20		**Evaluatie Mondiale dimensie in Leren voor Duurzaamheid**		**DSI**
21		**Evaluatie van het NCDO programma voor Sport en Ontwikkelingssamenwerking**	**April 2002**	**DSI**
22	**Art. 11**	**Evaluatie Financieringsmodaliteit op basis van outputfinanciering Koninklijk Instituut voor de Tropen**	**Aug. 2002**	**DCO**
23		**Evaluatie van de basissubsidie aan het Radio Nederland Training Centre 1997-2001**	**May 2002**	**DCO**
24	Art. 12	Het IntEnt programma, evaluatie over de jaren 1997-2001	April 2002	DDE
25		Evaluation and Strategic Review of the Consultative Group to Assist the Poorest	March 2002	DDE

It is difficult to give a precise picture of the financial significance and periods covered by the evaluated activities. In the majority of the evaluation reports the financial details supplied are either incomplete or non-existent, the reference period is not stated or budget figures only are given. Evaluation reports relating to projects financed by multi-donors generally omit any information on the contributions made by various donors and the recipient country. For practical reasons it was decided not to undertake a potentially time-consuming arithmetically-accurate reconstruction of commitments and actual expenditure. It is clear, however, that there is great variation in the financial significance of the evaluated activities.

In practice the Terms of Reference of the evaluations give a varied picture as regards the areas to be covered. Most are not organised in the way suggested in the Evaluation Guide in the Operational Procedures Manual (HBBZ), typically being lists of individual questions and areas of attention that relate mostly to progress of the activity and achievement of operational objectives at the level of implementing organisations (output). In some cases the normal progress reports on the activity being evaluated are so poor that getting information on progress becomes a major focus of the evaluation.

75% of the evaluations were formulated and organised by BZ itself. The remaining evaluations were organised in a multi-donor context or wider partnerships. With one exception (nos. 1) all the evaluations were carried out by external experts. The majority of the evaluation teams consisted of 2 to 4 experts who were selected on an individual or corporate basis by the responsible BZ budget holder or multi-donor steering committee.

The methods for selecting consultants and awarding contracts varied. Of the evaluations examined by the field study, there were two cases where the contract was awarded by means of an internationally-published call for tenders. The other eight evaluation contracts were awarded by private tender or by direct contracting. The most important background documents, the Terms of Reference and consultant contracts, are almost always available in a paper file. The transparency of the selection procedure leaves much to be desired. None of the selected evaluations contained any formal report or written record of the selection procedure or reasons for the choice.

In a good three-quarters of cases it was possible to determine the cost of the evaluation. In these cases there was found to be little or no correlation between the financial significance of the activity evaluated and the cost of the evaluation. Most of the Terms of Reference assumed the evaluation would be completed in a period of between 4 and 6 weeks. However, the field study showed that this was generally over-optimistic; in practice it had proved difficult to complete evaluations fully in less than three months.

The field study also showed that the draft reports were almost always discussed with the parties directly concerned (in the main, the budget holder and the project/programme implementer). The transparency of the final phase is often unsatisfactory. Draft reports that were submitted are no longer to be found in the dossiers and only occasionally are comments and criticisms by the concerned parties documented in writing. Generally speaking, the evaluations assessed had proceeded harmoniously, the interested parties being in broad agreement with the investigators' findings and the form and content of the reports.

3.2 The quality of the evaluations

The following sections summarise the results of the quality assessment of the evaluations. The assessment was based on the quality indicators listed in Annex 3. To be able to assess quality it is important that the indicators are formulated as precisely as possible. In some cases vaguely-worded indicators had to be translated into usable operational terms. This process of operationalisation is explained in Annex 3.

The summary in the following table shows the 16 selected indicators divided into three groups, corresponding to the three quality criteria set out in the RPE.

Table 4 Assessment criteria for decentralised evaluations carried out in 2002

VALIDITY	
1.1	description of the reason for and objectives of the evaluation
1.2	placing of the subject of the evaluation in its policy and institutional context
1.3	description of the subject of the evaluation
1.4	formulation of the problem definition and of the questions the evaluation seeks to answer
1.5	soundness of the method of investigation
1.6	soundness of the analysis
1.7	consistency of the evaluation
RELIABILITY	
2.1	representativeness of the evaluation
2.2	recording of information sources
2.3	verification of information sources
2.4	smooth operation of the evaluation process
2.5	independence of the evaluation team
2.6	monitoring of quality
UTILITY	
3.1	relevance of the results of the evaluation to its intended use
3.2	completeness of the evaluation
3.3	presentation of the evaluation results

Each indicator was rated according to a four-point scale (Very Good, Good, Fair, Poor).
'Very Good' – all the elements of that indicator were covered by the evaluation.
'Good' – most elements were covered.
'Fair' – only a minority of elements were covered.
'Poor' – most elements were not dealt with clearly or not dealt with at all.

The results of the assessment of the quality of evaluations, together with the associated explanatory notes, are attached in Part 2 and are summarised in the following table.

Table 5 Assessment of the quality of decentralised evaluations – score table

N.	VALIDITY							RELIABILITY						UTILITY			Tot.
	1.1	1.2	1.3	1.4	1.5	1.6	1.7	2.1	2.2	2.3	2.4	2.5	2.6	3.1	3.2	3.3	VG/G
1	VG	VG	p	f	F	F	G	f	f	p	VG	p	p	VG	f	f	5
2	VG	VG	f	f	VG	F	G	G	VG	VG	VG	G	f	VG	G	f	11
3	G	G	f	f	F	F	G	f	f	f	G	p	f	VG	VG	f	6
4	VG	p	G	VG	VG	VG	G	G	VG	VG	VG	G	f	VG	VG	G	14
5	VG	VG	G	VG	VG	VG	VG	VG	VG	f	VG	G	f	VG	VG	G	14
6	VG	VG	G	f	P	P	f	f	VG	f	VG	G	f	VG	f	G	8
7	VG	f	p	f	F	G	f	f	f	f	VG	G	f	G	f	p	5
8	VG	G	f	VG	F	G	VG	VG	VG	f	VG	G	f	VG	G	f	11
9	VG	VG	G	VG	VG	VG	VG	VG	VG	G	VG	G	f	VG	VG	G	15
10	VG	f	f	f	VG	G	G	VG	VG	G	VG	G	f	G	G	VG	12
11	VG	f	p	G	P	G	VG	p	p	f	VG	G	f	VG	G	f	8
12	f	p	p	G	F	P	f	p	p	p	VG	G	f	VG	f	p	4
13	VG	G	f	f	VG	VG	VG	VG	VG	VG	VG	f	f	VG	f	VG	11
14	VG	G	f	VG	VG	VG	f	f	f	f	G	G	f	VG	VG	VG	10
15	VG	f	f	VG	VG	VG	VG	VG	VG	VG	VG	G	f	VG	f	VG	12
16	VG	f	f	f	F	F	VG	f	VG	f	VG	G	f	VG	f	VG	7
17	VG	f	p	VG	G	G	G	p	VG	f	G	G	f	VG	f	VG	10
18	f	p	p	f	F	P	f	p	VG	f	VG	G	f	f	p	p	3
19	VG	p	VG	f	G	P	f	f	VG	p	VG	G	f	VG	f	VG	7
20	VG	p	p	f	P	VG	VG	p	p	f	VG	G	f	VG	f	G	7
21	VG	VG	VG	VG	VG	F	VG	f	VG	f	VG	G	f	VG	f	VG	11
22	VG	f	G	VG	VG	G	G	VG	VG	G	VG	G	f	VG	f	VG	13
23	f	p	f	f	G	F	G	p	VG	p	VG	G	f	VG	p	p	6
24	VG	G	G	VG	VG	VG	G	G	VG	VG	VG	G	f	VG	VG	G	15
25	VG	VG	VG	VG	VG	VG	VG	G	VG	VG	VG	G	f	VG	VG	VG	15
VG/G	22	12	9	13	15	15	19	11	18	9	25	22	0	24	11	16	

The table shows that six evaluations score either Very Good or Good on almost all aspects (with VG/G scores of 13/15). They are:

(4) Programa Estrategias para la Consolidación y Fortelicimiento del Sistema de Parques Nacionales Naturales (Bogotá)

(5) Centre for the Study of Human Rights, evaluation of the 1997-2001 project (HMA Colombo)

(9) Evaluación del Fondo de Alivio a la Pobreza (La Paz)

(22) Evaluatie van de Financieringsmodaliteit op basis van outputfinanciering, Koninklijk Instituut voor de Tropen (DCO)

(24) Evaluatie van het IntEnt programma period 1997-2001 (DDE)

(25) Evaluation and Strategic Review of the Consultative Group to Assist the Poorest (DDE)

A group of eight evaluations score Very Good or Good on 9-12 indicators whilst the scores for the other 11 evaluations may be regarded as Fair or Poor on half or more of the indicators (with Very Good or Good scores varying from 3 to 8). The results are explained in further detail in the following sections.

3.3 Validity

For the purposes of the assessment the validity of the selected evaluations was rated on the basis of seven indicators. The results of the assessment are summarised in the following table.

Table 6 Validity of evaluations – score table

Indicators		Very Good	Good	Fair	Poor
1.1	description of the reason for and objectives of the evaluation	21	1	3	0
1.2	placing of the subject of the evaluation in its policy and institutional context	7	5	7	6
1.3	description of the subject of the evaluation	3	6	9	7
1.4	formulation of the problem definition and of the questions the evaluation seeks to answer	11	2	12	0
1.5	soundness of the method of investigation	12	3	7	3
1.6	soundness of the analysis	9	6	5	5
1.7	consistency of the evaluation	10	9	6	0
Total		73	32	49	21
(Percentage)		42	18	28	12

The outcome was not uniformly positive. Although five evaluations scored Good or Very Good on all indicators, other evaluations received Fair or Poor on one or more indicators. The results are commented on below.

Re 1.1 Description of the reason for and objectives of the evaluation

In most of the evaluations, the description of both the reason for and the purpose of the evaluation as stated in the Terms of Reference (and in the report) is clear. The most common reason is that an evaluation is required by the relevant implementation document or the financing agreement. In most cases the purpose is twofold: evaluation of the past and formulation of recommendations to provide for input for decision making on the future of the subject of evaluation.

Re 1.2 Placing of the subject of the evaluation in its policy and institutional context

The policy and institutional context is the totality of frameworks, structures and processes associated with the implementation of the subject in question. It is essential for this aspect to be included in the evaluation report if a Very Good rating is to be achieved. 12 evaluations score Good or Very Good on this point. In the other cases the description given of the policy and institutional context is either very brief or incomplete.

Particularly striking is the absence of any reference to the relevant objectives in the Ministry's Explanatory Memorandum to the Budget and the lack of attention devoted in many evaluations to cross-cutting policy themes such as the mainstreaming of poverty reduction and gender equality. The description of the institutional context is also frequently very brief, which means that it is unclear to what extent the achievements and effects can actually be attributed to the policy followed and/or the activity financed.

Re 1.3 Description of the subject of the evaluation

The description of the subject of the evaluation is unsatisfactory in many evaluation reports. Only three of the 25 reports score a Very Good rating. Few reports contain a separate section giving a clear and logic description of the activity, programme or project being evaluated. Some of the relevant details are often set out in the introduction but many aspects proper to this description are scattered throughout the report. Particularly conspicuous is the absence of any proper financial statement as to the funds actually spent. The budget is generally described very briefly and there is almost never any mention of the financial contribution of the partner(s). Similarly,

there is little or no mention of the organisational arrangements of the programme or project financed. Finally, there is often no description of the target group(s), or the description given is incomplete. This means that it is not clear which people, and how many, have been reached.

Re 1.4 Formulation of the objective of the evaluation/problem definition and of the questions the evaluation seeks to answer

Around half of the reports score Good or Very Good on this indicator. The rating is based on the presence of a proper formulation of the problem and of those questions the evaluation seeks to answer that relate directly to the effectiveness and efficiency of the subject to be evaluated. The issue of effectiveness is addressed by almost all reports, but in a wide variety of forms. Normally there is a direct question about effectiveness but the issue is sometimes also dealt with in the form of topics, investigative tasks or activities. The level of detail also varies, from very general to very detailed and with many sub-questions. The questions asked relate particularly to the achievement of operational programme/project objectives at the level of the implementing civil society and/or government organisations (the performance achieved) and much less to the effects on society and/or the organisations concerned.

In around three-quarters of the reports the efficiency of the financed activity is addressed either to a very limited extent or not at all. This applies to both efficiency of policy (cost-effectiveness) and efficiency of operational management.

In evaluations relating to bilateral development cooperation, in around half of the cases additional questions were asked about sustainability, effects on the environment and/or specific population groups (the poor, women) which relate to development cooperation policy objectives. These specific questions were not included in the quality assessment.

Re 1.5 Soundness of the method of investigation

The methodology used by the majority of evaluations is a combination of document review and a self-conducted field study based on interviews and, to a lesser extent, surveys or direct observation. In many reports, the methodology is dealt with in the introduction to the report, but the way in which it is dealt with and the level of detail vary considerably.

The description of the structure of the evaluation and the data-gathering methods was Good or Very Good in around half the evaluation reports. In these cases the explanation of the use of output and/or effect indicators, pre-defined or otherwise, was also Good or Very Good. In most cases the structure and methods are partly laid down in the Terms of Reference and partly based on the knowledge and experience of the evaluation team. Evaluations are often incomplete due to lack of time and resources, but this is rarely stated explicitly.

The soundness of the methods used in the remaining evaluations is rated Fair or Poor. This may be because the report is inadequate or because of methodological limitations, e.g. no field study. In these cases the evaluation criteria are also normally not properly applied or the output and effect indicators are not clearly specified.

Re 1.6 Soundness of the analysis

The main elements of an analysis that is rated as Very Good are that data and information have been analysed and interpreted systematically, the findings have been formulated clearly and are inferred from the analysis, any differences between intended and actual results are explained and gaps in the information identified.

The soundness of the effectiveness analysis is unsatisfactory in many of the evaluations, being impaired by a lack of clarity in the specification of output and/or effect indicators and/or the narrowness of the information base. Only a small number of evaluations use specific, quantifiable indicators; most are based on qualitative indicators that are often poorly founded or difficult to verify. In many cases the results achieved are not set out systematically and/or no clear distinction is made between different kinds of result (process, performance, effects). Rarely is there any explicit mention of the weakness of the information base or of gaps in the information available (e.g. absence of baseline information).

No account was taken of the efficiency analysis for rating purposes. In most cases the analysis is either highly unsatisfactory or completely absent. While it is true that many of the activities in question do not lend themselves to a quantitative approach (cost-benefit analysis), the lack of criteria and standards at both policy and operational management level means that not even a much less in-depth analysis was undertaken.

Re 1.7 Consistency of the evaluation

In general there is a Good or Very Good level of consistency between the questions for the evaluation, findings, conclusions and recommendations. No obvious inconsistencies or contradictions were found. Fair scores relate to cases where, e.g. a stage has been omitted (e.g. recommendations are based on findings without any intermediate conclusions or main findings having been formulated) or where the link is less apparent.

3.4 Reliability

The results of an evaluation are reliable if the outside world can count on the findings and results of the evaluation being correct. Six indicators were used in the assessment to measure the reliability of evaluations. These indicators and the results of the assessment are shown in the following table.

Table 7 Reliability of decentralised evaluations that were carried out in 2002 – score table

Indicators		Very Good	Good	Fair	Poor
2.1	representativeness of the evaluation	7	4	8	6
2.2	recording of information sources	18	0	4	3
2.3	verification of information sources	6	3	12	4
2.4	smooth operation of the evaluation process	22	3	0	0
2.5	independence of the evaluation team	0	22	1	2
2.6	quality control	0	0	24	1
Total		53	32	49	16
Percentage		35	21	33	11

The results are commented on below.

Re 2.1 Representativeness of the evaluation

The assessment of representativeness is based on the size and/or composition of the attached lists of persons or organisations interviewed or surveys performed and the supporting notes. Scores of Fair or Poor relate to evaluations where discussions were only held with the implementers/managers of the activity under evaluation and not with the ultimate target group(s), evaluations in which the representativeness cannot be assessed (because there are no explanatory notes, lists of persons interviewed or information on sources consulted) or where the sample was clearly too small. There is seldom any explicit mention of limitations as regards the representativeness of the evaluation and how such limitations were dealt with.

Re 2.2 Recording of information sources

The sources of information are given in 18 of the reports, although the extent to which they are given varies greatly, from a single mention in the text or appendices to systematic references throughout the main text or appendices. In the remaining reports lists of interviewees or documents consulted are either incomplete or omitted entirely.

Re 2.3 Verification of information sources

The purpose of verification is to check whether the information supplied by individuals or organisations accords with the facts. The RPE states that independent information sources should be used wherever possible. In practice, however, evaluators mainly limit themselves to information supplied by those directly involved (annual reports, monitoring data, minutes of meetings, internal evaluation reports). In most of the evaluations, verification of this information is a thorny issue. Although there is virtually always some form of triangulation, inasmuch as evaluators have spoken to a number of different individuals or groups, it is no easy task to determine to what extent this results in a true picture.

Scores of Good or Very Good were awarded to those evaluations (nine of the 25) where the evaluators had carried out their own field study outside the circle of project/programme implementers or managers directly involved. Scores of Fair or Poor were awarded to evaluations in which this type of study either did not happen or was limited in its extent, or where information from third parties was used without further verification or reference to its limitations.

Re 2.4 Smooth operation of the evaluation process

The reliability of the evaluation depends greatly on the extent to which the evaluation team can work freely and without disturbance and have access to all the available information. In four cases there were communication difficulties during the field study, travel delays or poor information, which interfered with the evaluation to some extent (but without the consequences being spelled out in the evaluation). In the other cases it is clear from reports and interviews that although evaluation teams had to work under severe time constraints it was nevertheless possible for the work itself to proceed smoothly, which meant that it was possible to perform the evaluation and form a judgement.

Re 2.5 Independence of the evaluation team

It is essential to the perceived credibility of evaluations that they are independent. Where evaluations are commissioned or financed by departments or missions, a completely independent evaluation is unfeasible. Apart from two evaluations, which were carried out by the responsible budget holder himself, all the evaluations were contracted out to external experts. Unless the evaluators were demonstrably dependent in any way on the interested partners, the beneficiaries or the activity being evaluated a score of Good was awarded. In one instance there was a relationship between the evaluator and one of the interested partners and a score of Fair was consequently awarded.

Re 2.6 Quality control

Assessment on this indicator is based on quality control by budget holders, the existence of evaluation data in an evaluation dossier, whether the evaluation results were put before the parties directly involved for their comment, and whether the results were commented on by independent experts. In the evaluations looked at during the field study there was no assessment of draft reports by budget holders against quality standards. Under the RPE rules it should be possible to determine the quality and independence of the structure, organisation and implementation of the evaluation from a dossier. In practice, this is possible only to a very limited extent. Although important background documents, the Terms of Reference and the evaluation report are usually held in paper files, information on the selection procedure and on the assessment and approval of the content of the evaluation reports is often sketchy or absent. In almost no case was there any qualification in the form of further information about discussions and follow-up arrangements.

Apart from tendering, there are other ways of increasing the independence and quality of the evaluations. One of them is to use an independent supervisory committee or other form of independent external supervision. In practice this form of quality control was not used. In a small number of cases a steering or supervisory committee was set up (formally or otherwise), but these are always bodies composed of the parties directly involved, in which the progress and/or results of the evaluation are discussed, rather than independent bodies.

3.5 Utility

The RPE states that the utility depends largely on the validity and reliability of the evaluation. The indicators used to assess the criterion of utility are additional to this. The basic assumption is that utility will increase in line with the extent to which (i) the findings, conclusions or recommendations are relevant to the evaluation's intended use, (ii) the evaluation actually answers the questions it set out to answer, and (iii) the results of the evaluation are presented clearly and accessibly.

The following table shows the results of assessment against the selected criteria.

Table 8 Utility of decentralised evaluations carried out in 2002 – score table

Indicators		Very Good	Good	Fair	Poor
3.1	relevance of the results of the evaluation to its intended use	22	2	1	0
3.2	completeness of the evaluation	7	4	12	2
3.3	presentation of the evaluation results	10	6	5	4
Total Percentage		39	12	18	6
		52	16	24	8

The scores are commented on below.

Re 3.1 Relevance of the results of the evaluation to its intended use

As stated earlier, in most reports and the Terms of Reference the description of the intended use of the evaluation was Very Good or Good. Without pronouncing on their accuracy or feasibility, the conclusions/recommendations of the evaluations are, broadly speaking and with just a few exceptions, operationally focused and geared to the intended use. There are no indications that timeliness of the evaluations was a problem.

Re 3.2 Completeness of the evaluation

The assessment was based on whether the evaluation questions set out in the Terms of Reference were answered. In 14 evaluations this was only partly so. As noted earlier (point 1.4) the points to be covered, as set out in the Terms of Reference, are limited and there are some aspects to which most evaluations make little or no reference, efficiency in particular. In the absence of any analysis of effectiveness and efficiency it is not possible to make any assessment of the policy followed. The completeness requirement of the RPE is therefore not met.

Re 3.3 Presentation of the evaluation results

In some cases the presentation of the evaluation results is unsatisfactory. 16 evaluation reports are graded Very Good or Good but in the remaining nine reports readability and/or accessibility is impaired by one or more of the following factors:
- management summary missing or of poor quality
- no account given of implementation of the evaluation
- no summary and no clear separation of findings and conclusions
- untidy layout (poor paragraphing and style)
- absence of any kind of illustration
- main text too long
- lessons not set out clearly

The assessment did not look in any depth at the use that was actually made of the evaluations because this is not in itself an indicator of quality. The assessment would need to be set up differently to gather the necessary information. The use made of evaluations was, however, looked at in the field studies. In most cases the activities evaluated were to be continued and the interviewees stated that the evaluation results had played a greater or lesser part in the decision on whether to continue the activity, although it is not possible to establish this in detail and with certainty.

3.6 Summary of findings

The evaluations that were assessed provide information mainly on the implementation of the activities and the achievement of operational targets at the level of the implementing organisations (performance). They give much less information about the effects (effectiveness) and efficiency.[3] Three-quarters of the evaluations fail to address the issue of efficiency or devote very little attention to it. Evaluations are largely structured from a managerial and future-oriented perspective

[3] Note that activity objectives often make no distinction between performance-oriented operational objectives and effect-oriented objectives.

and in most cases an assumption is made that the programme or project in question will continue. They are organised much less from a policy perspective, i.e. with a view to identifying the effects of policy, which demands a more in-depth analysis. Nor do the evaluations form part of a broader effort by the Ministry to systematically assess effectiveness and efficiency and use the information so obtained for budget and annual report purposes.
The picture of the quality of the evaluations is a mixed one. Six of them are rated Good or Very Good on almost all points. There is a middle group of eight evaluations in which these higher ratings were awarded on nine to twelve indicators, while the remaining eleven, i.e. slightly less than half, may be regarded as Fair or Poor.

Indicators on which a large number of evaluations received a Good or Very Good rating are: the description of the reasons for and objectives of the evaluation, consistency between the questions, findings, conclusions and recommendations, smooth operation of the assessment work and the correspondence between the intended use of the evaluation and the evaluation results. Many common shortcomings were identified in all three areas (validity, reliability and utility).

The most significant problems relating to **validity** are:

I. *failure to place the subject of the evaluation clearly in its policy and institutional context.* Determining whether BZ policy objectives had been met does not appear to have been regarded as a primary objective for any of the parties concerned, in spite of the importance that is attached to specific policy principles such as (for development cooperation): the focus on poverty, the target group perspective, sustainability, gender equality, etc. As an obvious consequence little consideration is given to the institutional context within which the policy intentions need to be achieved;

II. *inadequate description of the activity under evaluation.* Information, particularly about programme/project inputs and how the funding was spent, is often incomplete or even absent. One of the consequences is that the efficiency criteria is scarcely touched on;

III. *the soundness of the analysis* is particularly impaired by a failure to specify output indicators and/or effect indicators clearly, or to distinguish adequately between different types of results (process, output, effects), and by the narrowness of the information base (no baseline information). The quality of the efficiency analysis is highly unsatisfactory.

The **reliability** is impaired by:

I. *limitation of the depth* of the investigation to the project/programme managers and/or the small size of the sample. In most cases there is no explanation of the limitations that this places on the representativeness of the evaluation or how these limitations were dealt with;

II. the lack of attention generally given by the reports to the *quality of the information* on which the evaluation is based. Independent sources are rarely used. Although there is virtually always some form of verification and some form of triangulation check, inasmuch as the inspectors have spoken to a number of parties concerned or have drawn on various sources of information, it is generally not clear to what extent this results in a true picture;

III. the absence of any quality assurance *system* to check that the quality of the inspection work is satisfactory.

The findings relating to **utility** are that:

I. the conclusions and recommendations presented are usually relevant to the intended use;

II. in many evaluations the evaluation questions are only partially answered;

III. for a considerable number of the evaluations the *presentation* of the evaluation results is unsatisfactory.

3.7 Factors that affect quality

As part of the field study, documents were studied and the views of many parties who had an interest in the evaluations were analysed in order to identify factors that have a positive or negative effect on the quality of the evaluations. Factors affecting quality can be divided into methodological, policy and process factors.

At this stage there are methodological difficulties in applying a single, RPE-based framework for assessing the quality of evaluations. The methodological framework used in this assessment is aimed at *ex post* evaluations. However, in practice evaluations were found to have specific goals and to take place at specific times. In most cases there were special circumstances that led to the formulation of specific Terms of Reference and evaluation questions to which the budget holder wanted an answer at the time. In contrast with *ex post* evaluations, which are aimed at assessing the effectiveness and efficiency of the policy followed, some of the evaluations selected have more of the character of progress inspections, in which the questions relate mainly to the management, progress or possible continuation of the activity in question. In other cases there has been a conscious choice to limit the scope of the evaluation because: (i) elements of an *ex post* evaluation have already been covered by other investigations (e.g. a separate 'value for money audit' or impact assessment); (ii) activities were being phased out anyway and were only of limited relevance to policy; (iii) the available information provided a sufficiently clear insight and/or there was doubt about the added value of a costly, detailed study of target groups and efficiency. In all these instances the evaluations relate to specific moments during the implementation of programmes/projects. The fact that they are less focused on analysing effects and efficiency means that when the RPE assessment framework is applied to them, a number of aspects (in particular indicators 1.4 and 3.2) receive lower ratings.

Changes in policy priorities during implementation of activities and, in some cases, pressure on spending are two factors that contribute to the predominantly future-oriented perspective on evaluations. For example, the changeover from a project-based approach to a sector-wide approach to development cooperation meant that less importance came to be attached to planned project evaluations. The Terms of Reference of 23 of the evaluations contain future-oriented questions from which it can be deduced that, when the Terms of Reference were drawn up, there was already an intention on the part of the budget holder to continue the activity.

In many evaluation reports there is more emphasis on learning lessons for the future than accounting for the past. This raises the question: How far has the intention to continue influenced the tenor of the findings? The assessment was unable to establish any definite answer. This future-oriented perspective means that the staff

concerned regard evaluation as a different kind of instrument from the *ex post* evaluation instrument defined in the RPE.

In many cases policy officers and others, such as controllers and managers, are not entirely familiar with the principles on which evaluation is based. Some of them are aware of the existence of the Evaluation Guide in the Operational Procedures Manual and just one or two are acquainted with the RPE. Some controllers indicated that they were unclear as to their role in regard to the quality of evaluations and in particular to effectiveness evaluations. There is no quality-standards-based assessment of Terms of Reference and draft reports. Little use is made of the non-mandatory Evaluation Guide from the Operational Procedures Manual. The operational management quality assessment checklist states only in the most general terms that an effective monitoring and evaluation system must be in place, that relevant regulations must be complied with in that regard, and that significant recommendations must be followed. The recruitment of consultants is not organised in a professional manner.

The implementers of the evaluations complain almost unanimously about vagueness, lack of clarity and sometimes internal contradictions in the Terms of Reference, which have a negative effect on the quality of their work. Although the time of the evaluations is known well in advance, Terms of Reference are often drawn up and evaluators selected with not much time to spare. The selection criteria for consultants attach more significance to their knowledge of the sector or of specific subjects than to their evaluation expertise. Over the years a practice has developed of conducting short evaluations which, with a few exceptions, last from 4 to 6 weeks, irrespective of the type or complexity of the programme or project to be evaluated. As there are often no predetermined output indicators and effect indicators or usable internal monitoring data, evaluation teams must then work under great time pressure. The result is rushed work and a lack of time for analysis and reflection, which in some cases contributes to a superficial, careless and incomplete presentation of the evaluation results.

The assessment looked at the evaluators' professionalism, on the basis of the knowledge and experience set out in their curricula vitae, as a possible factor affecting quality. The outcome was that no connection was established between the professionalism of the evaluators and the quality of the evaluations. All the curricula vitae that were examined scored Good or Very Good on professionalism.

The effects of proper structuring and supervision of evaluations are most apparent in the evaluations organised in the multi-donor context, which score proportionately higher than the bilateral evaluations. The multi-donor evaluations were characterised by more tightly-drawn Terms of Reference, a relatively large evaluation budget, careful selection of consultants with international experience, intensive monitoring and supervision of the evaluation and wide dissemination of the results.

The predominantly future-oriented perspective affects the value attached to the evaluations by the parties directly concerned. In some cases evaluations which were awarded low scores in the quality assessment were rated positively by the parties concerned – despite acknowledging their shortcomings – because they were satisfied with the conclusions and recommendations and because the evaluation had achieved its intended purpose. In those cases the parties concerned, in forming their opinions, took little or no account of the quality of the investigation on which the conclusions and recommendations were based.

ANNEXES

ANNEX 1: TERMS OF REFERENCE FOR THE QUALITY ASSESSMENT OF BZ EVALUATIONS COMPLETED IN 2002

1. Reason for the assessment

There are now national quality standards for the tools used to evaluate the Netherlands Ministry of Foreign Affairs' (BZ) budget and accounts. These standards are set out in the Order on Performance Data and Evaluations in Central Government (RPE), which has been in force since 2002.

The Policy and Operations Evaluation Department (IOB) is charged with assessing the quality of evaluations carried out by the Ministry's departments and missions. With effect from 2003 IOB will carry out an annual assessment of the quality of a selection of the evaluations completed during the previous year. The main objective is to get an idea of the extent to which BZ evaluations meet the quality standards, and to identify factors that affect the quality of the evaluations and how, with the intention of using this information to improve the quality of evaluations. The results of the assessment are also important for supervision of the evaluation function by the Audit Committee and BZ's Financial and Economic Affairs Department (FEZ).

2. Background

The whole of the government service is making a concerted effort to make systematic use of evaluation tools as an integral part of the desired result-oriented management model, thereby implementing the provision in the Government Accounts Act which states that Ministers are responsible for carrying out periodic checks on the effectiveness and efficiency of policy. The RPE contains detailed regulations issued under the terms of the Act.

Evaluations relating to BZ are divided into:

- evaluations carried out by IOB, which focus primarily on specific policy themes, instruments or programmes. These are termed 'central evaluations';
- evaluations by the Ministry's departments or missions. These evaluations relate mainly to specific activities (primarily projects and programmes) and, to a lesser extent, themes and sectors. These are termed 'decentralised evaluations'.

In the 1990s guidelines for the quality of BZ evaluations were set out in the operational procedures manuals that were current at the time. Now, however, quality standards apply to the whole civil service and are laid down formally in the RPE. These standards are: validity of the design of the evaluation and its conclusions, reliability of the investigative methods used and utility of the results.

In April 2002 the note 'BZ Policy on Organisation of the Evaluation Function: Putting the Order on Performance Data and Evaluations into effect' was adopted by the Ministry's senior management. The note sets out the responsibilities for the BZ evaluation function. The relevant director or director-general determines the optimum and relevant level of coverage per policy article, while departments and missions are responsible for planning and implementation. FEZ is responsible for co-ordinating and promoting evaluation work. Under the Government Accounts Act FEZ must also report annually to the Netherlands Court of Audit and the Finance Ministry on the evaluations of the effectiveness and efficiency of policy. To this end, FEZ draws up BZ's budget-related evaluation programme each year. The recently established Audit Committee is responsible for central control of BZ evaluations and,

in that context, for making random checks to verify the periodicity, quality and use or potential use of evaluations. The memorandum confirms the IOB's role in assessing the quality of evaluations.

3. Objectives and assessment questions

The objectives of the quality assessment are to get insight into the extent to which evaluations completed in 2002 meet the RPE quality standards and to identify the factors that affected the quality of the evaluations and how, with the intention that this information can be used to improve the quality of evaluations.

The key questions the assessment poses are:

1. What happened in the area of decentralised evaluations in 2002?
The assessment will seek to identify what kinds of evaluations can be distinguished (theme, sector, programme, project, organisation, process, etc.); the subject to which they relate (description, BZ's financial contribution, period covered); the type of evaluation (final evaluation, interim evaluation, evaluation/formulation, etc.); the commissioning body (budget holder, joint, multi-donor); who carried out the evaluation (internal or external implementers); cost and BZ's contribution thereto; and time taken to complete the evaluations.

2. To what extent do the evaluations completed in 2002 comply with RPE quality standards? The central question here is: To what extent are the evaluations valid, reliable and useful? A further point that will be looked at is whether evaluations have a policy function – i.e. whether any link is made with policy objectives and whether evaluation results feed back into policy.

3. What factors affected the quality of the evaluations and how?
The assessment will examine which factors have influenced the quality of the evaluations, either positively or negatively, including factors in the evaluation process. These will include factors in the decision-making process that led to the evaluation being carried out, in the formulation of the Terms of Reference, in the selection of evaluators, in management of the evaluation, possible refinements in structure, in methods and planning after the Terms of Reference have been drawn up, in implementation of the evaluation, and in presentation and distribution of results.

4. Assessment approach and methods

4.1 Scope of the assessment

The assessment comprises a desk study of a selection of 25 of the evaluations completed in 2002 and a supplementary field study of a selection of 10 of those 25.

In order to make a careful selection of evaluations for the quality assessment, a number of basic details from all 83 completed evaluations were analysed, i.e. the relevant policy article, the department or mission responsible, the type of subject being evaluated (project, programme, sector, theme, organisation, process, other), the type of evaluation (final evaluation, interim evaluation, review, evaluation/formulation, other) and commissioning body (internal, joint, multi-donor).

To obtain a broad picture of the quality of evaluations and the factors that contribute to it, the following criteria were adopted for the purpose of selecting evaluations for assessment:

- a spread of policy articles;
- a spread of departments and missions;
- a spread of different types of evaluation selected according to subject, time and organisation of the evaluation (project/programme, sector, theme, interim evaluation, final evaluation, evaluation by third party, joint evaluation).

Ten evaluations, by the DSI and DCO departments and the missions in Maputo and Nairobi, were selected for both desk study and field study.

Although the assessment is not completely representative, it nevertheless covers a fairly broad spectrum of the completed evaluations. Because of the diversity of the evaluations a representative assessment would have required a very large sample. This option was discounted, given the limited scope of many of the evaluations in comparison with the likely cost and added value of the quality assessment.

4.2 Assessment methods

The assessment comprises a desk study and field studies that will result in a final report containing the individual studies.

Desk study:
The desk study will look at the characteristics of the 25 selected evaluations and analyse and rate the evaluations, based on the Terms of Reference and the relevant evaluation reports. A list of characteristics, criteria and indicators based on the RPE quality standards will be used for this purpose. Any information on the evaluation characteristics listed that is lacking will be obtained from departments and missions. Ratings will be awarded on a four-point scale on the basis, wherever possible, of the Terms of Reference and the evaluation report. Where ratings cannot be traced back to the list of criteria and indicators, an explanation will be given. Gaps in the information available about evaluations and any assumptions about the quality of evaluation will be identified for further investigation by means of a field study of 10 of the evaluations.

Field studies:
The object of the field study is to check the findings of the desk study, fill in gaps in the information and verify assumptions about the quality of evaluations. The field studies will have a particular role in gathering information that could provide an insight into those factors in the evaluation process that affect the quality of evaluations, and how they affect it. This information will be gathered by examining relevant locally-available files on the activity being evaluated and on the evaluation itself, as well as by conducting interviews. Interviewees will include the embassy staff involved, representatives of the government of the country concerned, other donors, persons commissioning the evaluations, evaluators, members of any steering or supervision committees, those responsible for the activity to be evaluated and other interested parties. Specific checklists will be drawn up for these interviews.

5. Organisation of the assessment

5.1 Role of those with a direct interest

To ensure that the assessment is of good quality and that the results give rise to appropriate action it is essential that both those with a direct interest and potential users of the assessment at BZ play an active role by contributing to and

commenting on the Terms of Reference, implementation of the assessment and its results. Such assistance will not affect the fact that ultimate responsibility for the assessment lies with IOB.

The parties within BZ that have a direct interest are considered to be: the departments and missions responsible for the evaluations, DGIS, FEZ and the Audit Committee. These parties will be involved in the assessment in the following way:

- when the assessment commences they will be given an outline of what is planned;
- the draft Terms of Reference for the assessment will be provided to the departments and missions responsible for the selected evaluations and to FEZ and DGIS;
- as part of the desk study each individual rating of an evaluation will be submitted to the responsible mission or department for comment;
- the findings of the field studies will be submitted to the responsible department or mission for comment;
- the part of the report relating to the desk study will be discussed with FEZ and DGIS;
- the draft final report will be sent to departments and missions concerned for comment;
- the final report will be discussed with FEZ, the Audit Committee and DGIS.

5.2 Supervision and implementation

IOB inspector Ms R. Tesselaar will supervise the assessment and take part in the field studies. She will be involved in drawing up the final report and will have ultimate responsibility for the assessment. Two IOB inspectors, Mr D.C. van der Hoek and Ms A. Slob, will be involved in the assessment as co-readers. The task of the co-reader is to advise on the structure, form and implementation of the assessment and ensure that the documents produced include comments.

An experienced evaluator will be taken on to implement the desk study and field study. He/she must have intimate knowledge of the RPE quality standards and experience in the field of foreign policy, particularly international co-operation. He/she will carry out the desk study and will present the results in a sub-report. The ratings awarded to the evaluations will be determined in association with the inspector responsible, to ensure that they have been considered thoroughly and are consistent. The field studies will be carried out jointly by the evaluator and the inspector responsible.

5.3 Reporting

A final report of the results of the assessment will be published, based on the desk study and field studies. An abridged version of the final report will be translated into English.

5.4 Timetable

The assessment is expected to take six months, commencing on the date of approval of the Terms of Reference. This is based on the assumption that the desk study and field studies will each take approximately two months and run consecutively. The other two months will be devoted to preparing the field studies, drawing up the final report, taking account of comments and approving reports.

ANNEX 2 QUALITY ASSESSMENT OF DECENTRALISED EVALUATIONS: METHODOLOGICAL FRAMEWORK

A. General

In response to reports by the Court of Audit an Order on Performance Data and Evaluations in Central Government, known as RPE, was developed by an interministerial working group led by the Ministry of Finance. The Order has been in force since 1 January 2002 and contains detailed regulations under the Government Accounts Act. It has two linked objectives:

(i) to ensure that evaluations carried out in central government are sufficiently robust;
(ii) to ensure that policy information used in the Ministry's budget and annual report meets the relevant quality standards for this type of information.

Evaluations are defined in the RPE as systematic study in which, either during implementation of policy or subsequently, policy and/or operational management are evaluated against the criteria of achievement of objectives, effectiveness and efficiency. Evaluations can be broken down into:

(i) achievement of objectives – examining the extent to which the intended effects of policy objectives have been realised;
(ii) effectiveness – examining the extent to which the intended effects were realised as a result of the policy pursued;
(iii) an examination of the efficiency of policy and the relationship between cost and the effects achieved;
(iv) efficiency of operational management – examining the cost and quality of products/services delivered or output achieved.

The main objectives of the current quality assessment of decentralised evaluations are to assess the quality of decentralised evaluations completed in 2002 against the quality standards, validity, reliability and utitlity, formulated in the RPE, and to gain insight into the factors that have affected the quality of those evaluations and how, with the intention of using this information to identify gaps, problems and areas for improvement. A list of indicators and possible ratings has been developed to assess the quality of the evaluations, based on the RPE standards.

B. Validity

'Validity' refers to the overall validity of an evaluation. According to the RPE it will meet the following essential quality standards:

- the issue to be addressed by the evaluation has been defined and/or evaluation objectives identified; evaluation questions have been drafted;
- if the evaluation is put out to tender: the contract is consistent with the issue addressed by the evaluation, the objectives of the evaluation and the associated questions;
- the concepts to be evaluated have been operationalised correctly ('assess what you want to know'): the entities to be evaluated are defined consistently and are representative of the issue to be addressed, the objectives and the associated questions;
- a specific evaluation design and methods of investigation have been chosen and reasons given for the choice;
- the data obtained from the evaluation have been correctly analysed;

- there is a clear and consistent link between the questions, findings, conclusions and recommendations.

Whether and to what extent the selected decentralised evaluations meet these standards, will be assessed against the following indicators:

1.1 Description of the reason for and purpose of the evaluation:
- the reason for the evaluation has been described;
- the intended use of the evaluation has been formulated in clear terms.

1.2 Placing of the subject of the evaluation in its policy and institutional context:
- the relevant policy context has been set out;
- the institutional context of the activity to be evaluated has been described.

1.3 Description of the subject of the evaluation:
- long-term and short-term objectives have been described;
- the organisational arrangements and method of implementation have been described;
- the period covered has been specified;
- the resources used (inputs) have been stated;
- the target group of the activity has been identified;
- expected results (outputs, effects) have been described.

1.4 Formulation of the problem definition and evaluation questions:
- the activity evaluated has at least been examined and assessed for effectiveness (the extent to which intended effects were achieved as a result of the activity under evaluation) and efficiency (relationship between cost and results and/or between cost and the quality of the products/services/ changes delivered and output achieved).[4]

1.5 Soundness of the methods of investigation:
- the evaluation criteria used have been defined in accordance with accepted definitions (RPE, DAC);
- testable or measurable indicators have been used;
- methods used for data gathering have been stated and reasons given for choosing them;
- data have been gathered systematically.

1.6 Soundness of the analysis:
- data and information have been analysed and interpreted systematically;
- the findings have been formulated clearly, based on analysis and interpretation of available or collected data and/or information, and are thus firmly founded;
- gaps in the information available have been reported;
- differences between intended results and those actually achieved have been explained.

1.7 Consistency of the evaluation:
- there is a clear and consistent link between the questions, findings, conclusions and recommendations.

[4] Where no data are available on achievement of objectives, this criterion has generally still been assessed, either separately or as part of the effectiveness criterion.

C. Reliability

According to the RPE, evaluations must meet the following quality standards for reliability:

- the entities to be evaluated have been selected and collected in an appropriate manner (e.g. drawn at random from a sufficiently large sample);
- the characteristics of the entities evaluated are identified in a valid manner that is open to subsequent verification;
- independent data sources are used wherever possible; the independence and/or professionalism of the party (internal or external) actually implementing the evaluation must be guaranteed;
- integrity – the information obtained by the evaluation tools accords with reality (no information has been improperly withheld or lost);
- continuity – the systems used to produce regular performance data and for evaluations have operated without problems.

Whether and/or to what extent the selected decentralised evaluations meet these standards, will be assessed against the following indicators:

2.1 Representativeness of the evaluation:

- interested parties have been identified;
- all interested parties and/or identified target groups – or a sample of them – have been listened to and consulted;
- the sampling and assessment methods have been documented and are acceptable.

2.2 Recording of information sources:

- the sources of the information used have been stated;
- the activities of the evaluation team have been stated;
- a list of persons interviewed and documents consulted has been appended.

2.3 Verification of information sources:

- the report deals with the question of reliability and/or independence of the information sources used and identifies possible problems in this area.

2.4 Smooth operation of the evaluation process:

- the evaluation assignment can be implemented within the allotted time and resources;
- the evaluation process was trouble-free.

2.5 Independence of the evaluation team:

- the evaluation team is not under the control of the person with responsibility for the management, design or implementation of the activity being evaluated;
- the team is independent of any partners in and beneficiaries of the subject of evaluation.

2.6 Overall monitoring of quality:

- draft evaluation reports have been checked for quality by budget holders;
- the information about the evaluation is available in an evaluation dossier;
- evaluation results have been submitted to all parties directly involved for their comments;
- evaluation results have been commented on by independent experts.

D. Utility

The RPE states that validity, reliability and accuracy are significant factors in determining the ultimate utility of the results. The following factors also tend to increase the utility of the results of evaluations:

- the intended use or purpose of the evaluation has been described and is geared to the information requirements of civil-service and political decision-makers (thereby improving the relevance of its content);
- the terms of reference and the questions posed by the evaluation ensue logically from the intended use;
- the results of the evaluation can be put to practical use; the conclusions and/or recommendations are consistent with the intended use;
- the results of the evaluation are presented clearly and in an accessible form
- the evaluation report includes a summary;
- the results of the evaluation are relevant: they are consistent with the intended use and, consequently, fulfil the information requirements of civil-service and political decision-makers;
- the results of the evaluation are available on time: they fit in with the budget/account-linked administrative and political decision-making cycles.

Whether and to what extent the selected decentralised evaluations meet these standards will be assessed against the following indicators:

3.1 Relevance of the results of the evaluation to its intended use:

- the results of the evaluation are relevant to the intended use of the evaluation;
- recommendations are policy-oriented and/or operationally-oriented;
- the results of the evaluation are available on time.

3.2 Completeness of the evaluation:

- the evaluation answers the questions raised in the Terms of Reference.

3.3 Presentation of the evaluation results:

- the evaluation report has a logical structure;
- the summary gives the essence of the evaluation;
- an account is given of the implementation of the evaluation (process, methods);
- findings, conclusions and recommendations are presented separately;
- the report is generally easy to read.

ANNEX 3 POINTS TO BE ADDRESSED BY THE FIELD STUDY

The field study is aimed at verifying the findings of the desk study and identifying gaps in information, especially factors affecting the quality of evaluations. The information will be gathered by studying relevant files and by conducting interviews with the staff concerned, with the persons who commissioned the evaluations, implementers/managers, evaluators, members of evaluation steering or supervisory committees and other interested parties. The questions to be answered by the assessment have been arranged into five groups.

A. Evaluation function, policy and systems

Evaluations are carried out against the background of the VBTB and the RPE, which emphasise evaluations as a tool for policy-objective-linked and result-oriented management. Specific evaluation policy has been developed for bilateral development co-operation. This policy emphasises the quality and relevance of evaluations, fulfilment of the specific requirements of DGIS and coverage of policy as laid down by the RPE. It expresses a preference for evaluations that are prepared and implemented in partnership with other donors and under the leadership of the Southern partner, and for increasing the evaluation capabilities of Southern partners.

Questions to be asked are:

- What does the budget holder think about evaluation policy and practice within BZ (function, systems, responsibilities, role of the Ministry in The Hague and of the mission)?
- Is the budget holder familiar with the RPE and the quality standards it sets out? What does the budget holder think about the quality standards set out in the RPE? Have they been applied? Are they useful?
- How does the budget holder deal with the requirement for independence of the evaluation team? Have measures been taken to ensure that evaluation teams are independent?
- In most cases the emphasis in evaluations is on learning operational lessons. Policy-related aspects, linking of the subject of the evaluation to policy objectives, feeding results back into policy – these are rarely given any explicit mention. Why is this?
- What is the availability of monitoring data or other evaluative information sources that are used for evaluations?
- Almost every evaluation has difficulty in dealing with the question of efficiency. What is the problem?
- Where applicable - what does the recipient country think about BZ's and/or other donors' practice in carrying out evaluations?
- How does this practice fit in with the relevant policy and legislation in that country?

B. Formulation of the Terms of Reference

The Terms of Reference (ToR) play an important role as they constitute the assignment for those who will implement the evaluation and provide the legitimisation for carrying out the evaluation. Questions to be asked are:

- Who took the decision to carry out the evaluation?
- Who, at the level of the Dutch budget holder, is responsible for the evaluation? Does the person concerned have any evaluation experience?

- Who drew up the ToR? To what extent were other interested parties besides the responsible budget holder involved? How significant was the role of the other interested parties?
- Is the person who drew up the ToR familiar with the RPE and the quality standards set out in it? Was the RPE used? Was it useful?
- The various ToRs have been found to differ considerably in format and content. Is the respondent familiar with the Terms of Reference framework given in the Evaluation Guide in the Operational Procedures Manual? Is it used and if not, why not? Is it useful?
- In practice it has been found that the ToR and/or the evaluation report fail to give a precise description (objectives, instruments, activities, amount of finance involved, input, output, etc.) and delineation of the item under evaluation. Why is this?
- In most cases the emphasis in evaluations is on learning operational lessons. Policy-related aspects, linking of the subject of the evaluation to policy objectives, feeding results back into policy – these are rarely given any explicit mention. Why is this?
- How did the budget holder calculate the estimated cost of the evaluation? In the opinion of the budget holder/interested party/implementer, were sufficient resources available?
- Were the Terms of Reference formally adopted by the budget holder? What is the procedure in the case of multi-donor programmes? Who decides what?
- What did the implementing consultant think of the Terms of Reference that were formulated? Was there any further consultation with the implementer?

C. Selection and engagement of the evaluation team

The selection and engagement of suitable team members is an important factor in the success and/or quality of the review or evaluation. In most cases the selection criteria used to choose the team members are either not given in the Terms of Reference or are set out only in the briefest terms. Possible selection criteria are: knowledge of evaluation methods, independence, knowledge of the sector, knowledge of the region or country, proficiency in the working language, communication skills, leadership abilities. Questions to be asked are:

- What criteria are considered crucial? Must all the criteria be applied to all individual members?
- Are there any differences of opinion between the budget holder and other interested parties about application of the criteria?
- Is a consultants database available to the budget holder/interested parties?
- What do the evaluators/implementers think of the independence of the evaluation? Were measures taken by the commissioning body and/or evaluator to guarantee independence?
- Where applicable, what does the budget holder think of participation of team members from the recipient country? Were the other interested parties involved in the choice and/or selection of the evaluation team? If not, why not? What do other interested parties think about the choice/selection?
- In most cases a decision was taken to select and engage individual experts; in a few cases consultants firms were preferred. On what was this choice based? To what extent were the formal rules on the procurement of services adhered to?
- What do the implementers themselves think about the selection procedure?

D. Planning and implementation of the evaluation

In a small number of cases the evaluation team reported problems in implementing the evaluation. The broad experience of the assessment was that little or no information was given about problems in evaluation reports.

- To what extent are the evaluators familiar with the RPE and the quality standards it sets out?
- How feasible and achievable did the timetable turn out to be?
- To what extent was it possible to use information from other evaluative sources, such as internal programme/project monitoring and evaluation systems, and how much use was made of such information? Was the reliability of that information verified?
- The Notes on Review and Evaluation in the Operational Procedures Manual refer to an evaluation plan drawn up jointly by the mission leader and the commissioning body/budget holder. Was such a plan drawn up? If not, why not?
- In a number of cases there were problems with the timing of the evaluation. What was the cause?
- How did the evaluation team members integrate? Was the workload distributed effectively? What part did the mission leader play in this?
- Was there a good relationship between the evaluation team and the commissioning body and other interested parties? Were there any differences of opinion? What were they about?
- In most cases the field study was sample-based. Who decided what the sample would be and how? In most cases the sample is poorly documented. What is the reason for this? How did the implementer verify the information sources?
- What steps were taken to increase the quality and independence of the evaluation (supervisory committee)?
- Were any specific measures taken to optimise the utility and actual use of the evaluation results?
- Is an evaluation dossier available? What are its contents?
- Was there sufficient opportunity for debriefing and discussion of the findings? What form did the debriefing take (oral, based on written summary, based on draft report)?

E. Reporting

- The desk study showed that the quality of reports was very variable, also within one division or mission. What is the reason for this?
- What does the budget holder/interested party think about the quality of the report? Does the report satisfy the requirements of readability and completeness? If not, what were the contributory factors (professionalism of the evaluation team, other factors)?
- What does the implementer him- or herself think about the quality of the report? What factors were involved?
- Has the framework for review/evaluation reports in the Evaluation Guide in the Operational Procedures Manual been used? If not, why not? If it was used, was it easy to use?
- The placing of the subject being evaluated in its policy and institutional context and the description of the subject leave much to be desired in many cases. What is the reason for this?
- It is preferable/best if the measurement of effectiveness is based on objectively verifiable and measurable indicators. The progress of an activity can be measured on the basis of input indicators (to determine the extent to which the

inputs were achieved), output indicators (extent to which outputs were achieved) or effect/impact indicators (extent to which the objectives were achieved). Effect/Impact indicators are seldom used. Why?
- Almost every evaluation has difficulty in dealing with the question of efficiency. What is the problem?
- Is there any specific assessment procedure for reports at the level of the budget holder/interested party? Does the budget holder/interested party have criteria for assessing the report (requirements as to form, quality)?
- Was the implementer notified of the comments made by the budget holder/interested parties? How (orally/in writing)? Were they in complete agreement? How were different points of view incorporated into or dealt with in the draft report and final report? Did they give rise to significant amendments?
- Was there any standard procedure at the level of budget holder/interested party for finalisation of the evaluation? Was the final report formally adopted by the budget holder?

F. Utility of the evaluation results

- How useful did the evaluation turn out to be? How is this apparent? Did the evaluation produce results that have actually been used in the work processes and/or policy? Were useful lessons identified?
- In a considerable number of cases, the general presentation and readability of the reports are unsatisfactory. What is the reason for this?
- Who was responsible for distributing the report and how was this done?
- What was the response of the budget holder/interested party to the findings, conclusions and recommendations of the report? How did they react to criticism?
- To what extent were the results of the evaluation reported to the policy departments concerned? Were parties without a direct interest (e.g. the international donor community) also informed?

ANNEX 4 DOCUMENTS CONSULTED

*Order on Performance Data and Evaluations in Central Government (RPE), The Hague, February 2001

*BZ-policy on the organisation of the evaluation function, translation of the Order on Performance Data and Evaluations in Central Government (RPE), 2002

*DGIS – Evaluation Policy Bilateral Development Co-operation, 2002

Evaluation and Monitoring, Netherlands Development Cooperation, Summary Evaluation Report 1995

"Evaluation programme 2001-2007, Ministry of Foreign Affairs, FEZ, May 2002

*Assessment of the evaluation function of the Ministry of Foreign Affairs – study on the quality of evaluation reports, R.A. van de Putte, W. Flikkema, April 2003

*BZ Operational Procedures Manual, The Hague, 2003

*Netherlands Court of Audit, State of Policy Evaluation, 2000

*Netherlands Court of Audit, Care for Policy Information, 2003

* documents that are available in the Dutch language

ANNEX 5 DEFINITIONS OF EVALUATION TERMINOLOGY

RPE definitions

Evaluation *ex post*
In the RPE *evaluation ex post* is defined as 'systematic examination of the effects of existing policy, of the way in which policy is implemented and/or of the cost and quality of products/services delivered'.

This type of evaluation is divided into a number of elements:
(i) achievement of objectives – examining the extent to which the intended effects of policy objectives have been achieved;
(ii) effectiveness – examining the extent to which the intended effects were realised as a result of the policy implemented;
(iii) an examination of the efficiency of policy and the relationship between cost and the effects achieved;
(iv) efficiency of operational management – examining the cost and quality of products/services delivered or output.

Final effects
The ultimate consequences of policy for society or government organisations.

Intermediate effects
The intermediate, manageable effects that contribute to achievement of the final effects.

Performance data
Concise, structured information about the effects of policy, what was done to achieve them and what it cost. This information can be obtained both from systems of regular performance data and by carrying out periodic evaluations.

HBBZ and/or IOB definitions (adapted to development co-operation)

Evaluation
The systematic and objective assessment of the relevance, effectiveness and efficiency of a current or completed project, programme or policy.

Review
The assessment of operational aspects and performance of a programme or project, periodically or on an ad hoc basis (programme/project set-up, implementation structure, progress of implementation, etc.)

Relevance
The extent to which the effects of activities that have been implemented contribute to the stated goal, the impact.

Effectiveness
The extent to which the direct results of activities, the output, contribute to the attainment of the programme objective(s), i.e. the outcome. Programme objectives are the objectives that the activities are intended to achieve.

Efficiency
The extent to which the results achieved (the output) outweigh the cost of the chosen resources (the input) and the way in which they were used.

Ministry of Foreign Affairs | P.O. Box 20061 | 2500 EB The Hague | The Netherlands

Policy and Operations Evaluation Department

ISBN 90-5328-346-3

ORDERCODE: BZDR0640/E

Ministerie van
Buitenlandse Zaken

BZDR0640/E