Sida

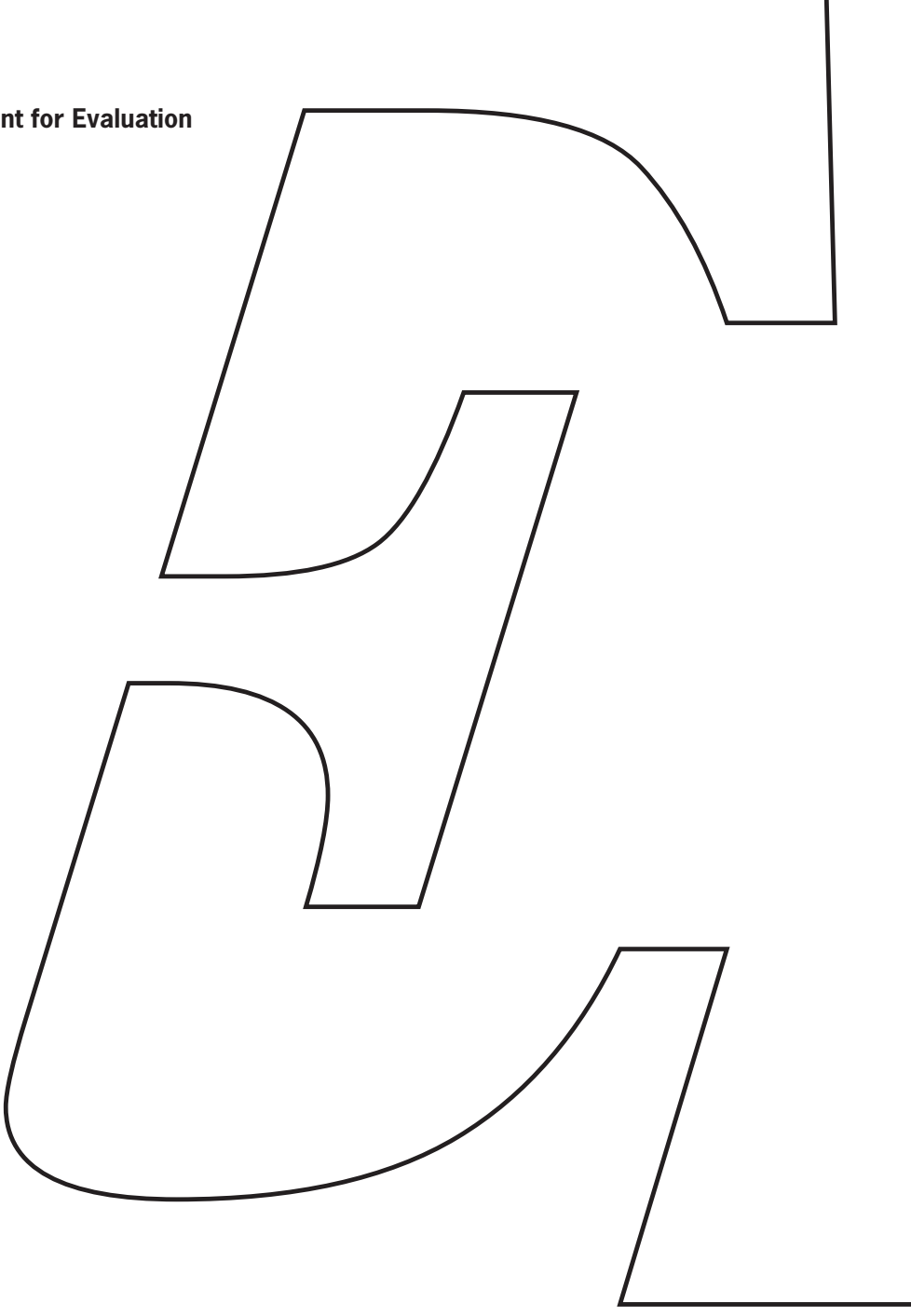
Are Sida Evaluations Good Enough?

An Assessment of 34 Evaluation Reports

Kim Forss, Evert Vedung, Stein Erik Kruse,
Agnes Mwaiselage, Anna Nilsdotter

Sida Studies in Evaluation 2008:1

Department for Evaluation



Are Sida Evaluations Good Enough?

An Assessment of 34 Evaluation Reports

Kim Forss, Evert Vedung, Stein Erik Kruse,
Agnes Mwaiselage, Anna Nilsdotter

Sida Studies in Evaluation 2008:1

This report is published in *Sida Studies in Evaluation*, a series comprising methodologically oriented studies commissioned by Sida. A second series, *Sida Evaluation*, covers evaluations of Swedish development co-operation. Both series are administered by the Department for Evaluation, an independent department reporting to Sida's Director General.

This publication can be downloaded/ordered from:
http://www.sida.se/publications

Author: Kim Forss, Evert Vedung, Stein Erik Kruse, Agnes Mwaiselage, Anna Nilsdotter.

Sida Studies in Evaluation 2008:1

Commissioned by Sida, Department for Evaluation

Registration No.: 2005-004489
Date of Final Report: May 2008
Printed by Edita Communication, Sweden 2008
Art.no. SIDA45265en

ISBN 978-91-586-8183-5
ISSN 1402-215X

SWEDISH INTERNATIONAL DEVELOPMENT COOPERATION AGENCY
Address: SE-105 25 Stockholm, Sweden. Office: Valhallavägen 199, Stockholm
Telephone: +46 (0)8-698 50 00. Telefax: +46 (0)8-20 88 64
E-mail: sida@sida.se.
Website: http://www.sida.se

Foreword

This is an assessment of the quality of evaluation reports commissioned by Sida's line departments and Swedish embassies in countries where Sweden and Sida are engaged in development co-operation. Based on a close reading of a sample of evaluation reports published in the Sida Evaluations series it looks at the coverage, credibility and usefulness of the results information generated through the decentralised part of Sida's evaluation system.

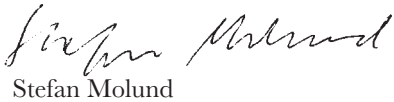
The purpose of the study is to contribute to on-going efforts by Sida's Department for Evaluation (UTV) and Sida as a whole to enhance the quality of Sida evaluations. Sida has recently adopted a programme for strengthening its system for results based management and evaluation is a key component of that system. The study will be very useful as a baseline against which to evaluate the effects of staff training programmes and other actions taken in order to improve the quality of Sida evaluations in years to come.

It should be noticed that the study was originally intended to be the initial step of a more comprehensive study that would also include a review of the actual use of the evaluation instrument in different country contexts. As a result of budget cuts and shortages of staff at the Department for Evaluation, however, the second part of the study had to be cancelled.

Notice also that the present study is an abbreviated and edited version of a considerably longer consultancy report originally delivered to Sida. One of the chapters of the original report is included as an annex. The study was abbreviated and edited for reasons of accessibility.

While UTV has been much involved in the editing of the report it is not responsible for the quality assessments that it contains. The latter belong entirely to the authors, a team of independent evaluators and evaluation specialists. The assessment process is described in the report.

According to the report, the quality of Sida evaluations is by and large not as good as it ought to be. The report is handed over to Sida with the expectation that it will generate a determined response.



Stefan Molund
Acting Director
Department for Evaluation

Table of Contents

Executive Summary

Introduction

Evaluations are 'reality tests' of aid efforts and strategies intended to be used in support of accountability, decision-making and learning. In development co-operation today, there is increased demand for evidence-based results information and greater emphasis on results-based management. The purpose of this study is to contribute to ongoing efforts by Sida's Department for Evaluation (UTV) and Sida as a whole to improve the quality of Sida evaluations.

The study is based on a close reading of 34 evaluation reports published in the Sida Evaluations series between 2003 and 2005. All the reports were produced by Sida's line departments and the Swedish embassies in countries where Sida is involved, and most of them focus on individual projects and programmes. UTV evaluations, which are usually concerned with wider issues, were deliberately excluded from the study.

The reports were analysed by an external team of evaluation specialists in order to find out whether the quality of the evaluations produced by the line departments and the embassies should be considered good enough. Do Sida evaluations produce information on processes and results that is comprehensive and detailed enough in view of Sida's management needs and reporting requirements? Are findings, conclusions and recommendations well supported by reported evidence? Do the evaluations produce lessons that are useful for learning and improvement beyond the evaluated projects and programmes?

The overall answer is that there is much room for improvement. Although there are exceptions, Sida evaluations are by and large not good enough. The study concludes with a series of general recommendations for improvement.

The Assessment

An evaluation, as a process, can be divided into four main phases: (1) the specification of a set of evaluation questions, (2) the search for answers to those questions, (3) the organisation of the answers into a report, written or verbal, and finally, (4) the use of the report for purposes such as management or learning. This study has focused on the first three phases in so far as they could be assessed from the reports.

It should be noted that this is a desk study and that it has nothing to say about the actual reception and use of the evaluation by its stakeholders. As use is an important quality criterion for evaluation processes, this is an important limitation. Nevertheless, while the study provides no information on the actual use of the evaluations, it has much to say about their potential usefulness.

The assessment focuses on the following issues:

- the quality of the Terms of Reference (TOR) for the evaluations and the extent to which the evaluation reports adequately responds to those TOR;
- the quality of the design of the evaluation, including its data collection methods;
- the quality of the information on results and implementation;
- the quality of conclusions, recommendations and lessons learned.

For each of these issues there was a set of quality criteria against which the reports could be systematically rated. The rating was done by the team of external evaluators and evaluation specialists who had also defined the criteria. Each of the reports was read by at least two of the team members and the results were discussed one report at a time in the wider group. The resulting assessments thus represent the reflected collective opinion of the rating team.

Findings

The findings are conveniently summarised as answers to a series of questions:

1. *Are the TOR for Sida evaluations well formulated and do the evaluations adequately address the evaluation questions formulated in the TOR?*

Most of the evaluations in the sample addressed the questions raised in the TOR, though they did not necessarily provide satisfactory answers (cf. below). As evaluation teams usually present draft reports to Sida and are asked to make adjustments, where necessary, it is not surprising that the end product corresponds fairly well to the TOR. The TOR were not always clearly formulated and focused, however. The overall assessment of the TOR for the evaluations examined in this study was not very good.

2. *Do Sida evaluations provide valid and reliable information on efficiency, effectiveness, impact, relevance and sustainability?*

Taking the limitations in time and resources into account, about two thirds of the evaluations contain a minimally satisfactory analysis of effectiveness, sustainability and relevance. Fewer than half, however, contain an adequate analysis of impact, and only one in five delivers a satisfactory discussion on

efficiency. While the majority of the reports (74%) were found to address the questions in the TOR, between 30% and 80% of Sida's evaluations fail to deliver plausible statements for each of the five evaluation criteria.

Most of the evaluations cover effectiveness appropriately (62%), although often in the sense of goal achievement at the output or near outcome stages. Many evaluations that draw conclusions for intervention effectiveness do not give the issue of attribution sufficient consideration, i.e. they do not show any empirical evidence of the *intervention* having an influence.

Impact studies are less common (47%), if we take "impact" to mean the effects of the interAvention itself as opposed to the effects of concurrent extraneous factors. Causal analysis should be an integral part of effectiveness and impact assessment. In the sample reports, the outcome objectives that are to be assessed are often broad, long-term and of a multiple nature. In many cases the evaluations are designed in a way that makes it difficult to assess the actual impact of an intervention (see question 3).

Most evaluations do not consider efficiency sufficiently: only 21% of the evaluations in the sample succeed in this task. Financial analysis is a weak area in most reports, and the cost of interventions is rarely analysed and compared to outcomes or impacts – not even at a general level. Questions about the extent to which more and better outcome effects might have been achieved by alternative means are rarely addressed. All too often, conclusions about efficiency are presented without empirical data to support them.

With regard to their assessment of sustainability, 59% of the evaluations are rated as satisfactory. Few evaluations apply the sustainability criterion well, however, and the analysis is often too impressionistic. In many cases, broader and more systematic analysis covering different aspect of sustainability would have been useful.

The assessments of relevance are found to be somewhat more accurate and adequate, though in most cases relevance is assessed in relation to Sida's and the respective partner country's policies. There is no systematic discussion of relevance with respect to the needs and priorities of the target group.

3. *Do Sida evaluations contain a clear and consistent analysis of attribution and explain how and why the interventions contributed to results?*

Very few evaluations contain a satisfactory analysis of attribution and causal mechanisms. The evaluations frequently present data bearing on the indicators set out in the logical framework of the intervention, but they do not adequately assess the extent to which the recorded changes can be explained by the intervention. Nor is the issue of unintended consequences addressed in most cases.

4. *Do Sida evaluations have an appropriate research design?*

The evaluation design is considered appropriate in the majority of the cases, given the constraints of time and resources. Nonetheless, 21% were rated as "not quite adequate" or as suffering from "significant problems". The most common research designs are narrative analysis (65%) and case studies (35%). None of the evaluations used experimental or quasi-experimental designs. Impact analysis would in many cases have required a stronger design to generate valid and reliable conclusions.

With regard to data collection methods, the assessment is less favourable. One in three evaluations was found to lack appropriate methods for answering the evaluation questions. Most evaluations rely on a basic mix of methods, with open-ended interviews and document analysis being the most common, sometimes combined with ad-hoc observations. Few evaluations use focus group interviews, structured interviews or surveys, and standardised interviews and structured observations are rare.

Sampling is usually purposive or purely ad hoc, with the evaluators tending to rely on the information that is most easily available. Only two evaluation reports contain any discussion of the principles they applied when selecting the sample and how this affected the findings.

5. *Is the evaluation process in Sida evaluations well documented and transparent, so that readers can make an independent assessment of validity and reliability?*

Fewer than two thirds of the evaluations contain an adequate section on methods and methodology, and even fewer discuss validity and reliability (35%) or the limitations of the task (41%). Most of the reports do not include their data collection instruments or present data to support their conclusions. This means that the reader often does not have a chance to make an independent assessment of the evaluation methodology. For an evaluation report to appear reliable it must explain how indicators are defined and data collected.

6. *Do Sida evaluations include a valid and reliable analysis of the management of interventions?*

An analysis of management aspects is not necessary or relevant to all evaluations. Nonetheless, many of the evaluations include an analysis of one or two dimensions of management, such as planning or organisational structures, while few contain a comprehensive assessment of implementation issues. Fewer than half provide a satisfactory analysis of organisational structures, co-ordination and networks, and fewer still include a sufficiently instructive analysis of leadership, planning and financial management. It is striking how leadership and governance issues are often left out or only marginally discussed.

7. *Do Sida evaluations provide clear and focused recommendations for specified target groups?*

The majority of evaluations have clear and consistent recommendations that are derived from the analysis and conclusions. As evaluations are often meant to be used for decision-making, it is valuable that most of the reports were found to deliver practical recommendations that could be translated into decisions for clearly specified groups of actors

As many of the evaluation reports do not have sufficient evidence to support their findings and conclusions (cf. above), however, the quality of the recommendations derived from those must be considered as questionable.

8. *Do Sida evaluations document interesting and useful lessons learned from the interventions that were evaluated?*

"Learning" is one of the main purposes of evaluation. The "lessons learned" section in an evaluation report is meant to present new insights that are relevant to a wider audience than the immediate stakeholders. Lessons learned are supposed to generalise and extend the findings from the intervention under study, either by considering it as an example of something more general or by connecting it to an ongoing discourse. This requires familiarity with both the international development debate and the discipline or sector under study and may not be possible or even necessary in all cases. The degree of generalisation may also vary from case to case.

For all that, it is surprising that only 26% of the evaluation reports contain a section on lessons learned, and it is a cause for concern that the sections that where available are so weak. Only four reports were found to make strong contributions to the understanding and knowledge of development cooperation.

Conclusion and Recommendations

It must be concluded that evaluation quality assurance should be improved at Sida. There is a need for more and better empirical evidence and systematic use of such information in a majority of the reviewed reports. It is of particular concern that so few of the evaluations included enough information on the methods used. This made it difficult to assess whether the conclusions were reliable and clearly derived from the data. Reliable conclusions are in essence the purpose of evaluations.

Some of the weaknesses in the individual reports stem from poor TOR, which could have been picked up during the inception phase. This means that they are largely the responsibility of the Sida staff involved in the management of evaluations. Other problems may be caused by a lack of technical skills or poor motivation among the consultants who carry out evaluations on behalf of Sida, and in many cases there seems to be a mismatch

between the questions in the TOR and the resources invested in answering them. A lack of recognition and reward for high-quality evaluation work appears to be yet another problem.

This report presents a multi-faceted picture of the quality problem, but no straightforward recommendation as to the approach to take in order to improve the quality of evaluation. There are quality issues at different levels and multiple strategies are required to improve quality:

1) Improving the quality of individual reports produced by external evaluators

Design issues need to be resolved in close cooperation between Sida and the consultants during the inception phase; more feedback could be given during the evaluation process; and increased use could be made of reference groups or other committees that can safeguard quality.

2) Assuring the quality of the evaluation system

Evaluation capacity within Sida needs to be strengthened and integrated into overall planning and management; sufficient financial and human resources for evaluation need to be secured; and communication of evaluation results should be improved.

3) Increasing the demand for and utilisation of evaluations

More attention needs to be paid to the timing and use of evaluations. Stakeholders – ranging from project managers to politicians – need to be provided with relevant information at the right time.

Given the increased focus on results-based management and the tendency of the general public and decision-makers to take evaluations at face value, as telling the truth, there is ample evidence in this report to suggest that more attention needs to be paid to the quality of evaluations at Sida.

1 Introduction

1.1 Purpose and Background

Swedish development cooperation has a history of more than 50 years, and evaluation has been a prominent part of the system for at least the past 40 years. In response to requests for reliable feedback from the Swedish Parliament, Government and Sida itself on the implementation and results of aid, Swedish and international consultants have produced hundreds, if not thousands, of reports. When Sweden takes part in international forums, there is often an emphasis on the need for high-quality evaluation systems and a call for improved effectiveness driven by evaluation and learning.

The present study is an assessment of the quality of a small sample of evaluations produced by Sida. It is based on a close reading of 34 recent reports from the *Sida Evaluations* series, which contains most of Sida's evaluation reports, and addresses questions concerning the scope, validity, and potential usefulness of the information generated by Sida's evaluation system as it currently operates. While dealing primarily with the quality of individual evaluation reports, it also reflects on the quality of the evaluation system as a whole. The practical purpose of the study is to contribute to ongoing efforts by Sida's Department for Evaluation to help strengthen Sida's evaluation system. As it is published at a time when Sida is engaged in a major review of its own organisation and attempts to focus more sharply on development outcomes, it provides a timely baseline assessment of strengths and weaknesses of a key component of Sida's existing system for results based management.[1]

The study was developed in close dialogue with Sida's Department for Evaluation (UTV) and initiated as an experiment in assessment methodology. The TOR were unusually brief, asking only for a description of the results information contained in the reviewed evaluation reports and an assessment of the quality of that information. The rest was left open for discussion.

While the analytical framework for the study was developed in close dialogue with UTV, the study itself and its evaluative contents belong entirely to its authors. UTV did not participate in the discussions on individual evaluation reports and had no hand in the quality ratings that emerged from those discussions.

[1] The position paper Strengthening Sida Management for Development Results presents Sida's approach to results based management in brief.

1.2 Scope and Limitations

As a process, an evaluation can be divided into four main phases:

1) The specification of a purpose such as management or learning and the identification of a set of evaluation questions matching that purpose,

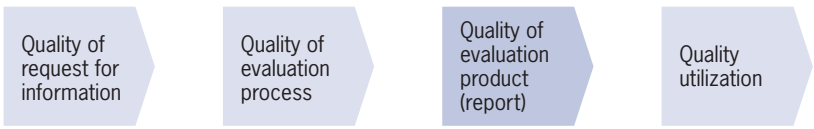
2) The search for answers to the evaluation questions,

3) The organisation of the answers into a report, written or verbal, and, finally,

4) The use of the report for its specified purpose.

As suggested in Figure 1 below, each phase of the evaluation process can be assessed in terms of quality. The evaluation questions set out in the TOR can be relevant, to a greater or lesser extent, to the specified purpose, as can the methodology to the evaluation questions. At each stage, steps are taken that are likely to affect the validity of the results and the usefulness of the final report.

Figure 1. Model of a systematic approach to evaluation quality



As this was a desk study, our information about the actual evaluation processes is limited. The conclusions are based on what is written in the final reports and on supplementary information about costs and other matters provided by Sida's Department for Evaluation (UTV).

This is an important limitation. While all the reports contain both the evaluation questions as they were first formulated in the TOR and the answers to those questions, other aspects of the evaluation process are not always well described. For example, the purpose of the evaluation is in many cases quite obscure, which means that the relevance of the evaluation questions is difficult to assess. The fact that the reports cannot tell us anything about how they were received and used after completion is obviously also a considerable limitation.

As we compiled the results of our assessments of the reports in the sample, we also reflected on the quality of the wider evaluation system producing them. We thus tried to assess the usefulness of the information contained in the reports for results analyses in the aggregate in much the same way as we sought to assess the usefulness, or potential usefulness, of individual evaluations for their particular stakeholders. For example, while noting that it might be quite in order for any particular evaluation not to raise questions about

the efficiency of the activities reviewed, the fact that questions about efficiency were usually not answered by Sida evaluations should perhaps be described as a weakness of the system as a whole.

Nonetheless, our assessments of quality at corporate level are tentative and limited in scope. Most importantly, we do not deal with processes of evaluation programming. As we do not know why certain activities were singled out for evaluation during the reviewed period while others were ignored, an assessment of the quality of the overall system is obviously beyond our purview.

1.3 Quality Criteria and Ratings

Our first step was to specify exactly what we meant by a good evaluation report. What are the different evaluative criteria to be used in assessing evaluation quality? It was agreed, for example, that a good report should provide answers to the questions in the TOR and be well structured, so that the reader can follow the arguments and find his or her way through the text. We also agreed that in a good evaluation report the conclusions should be reliable and clearly derived from the data. The report should, of course, also be well written.

Our criteria of what constitutes a "good" evaluation report were taken from literature on the subject. The OECD/DAC Trial Evaluation Quality Standards is a key document for assessing the quality of Sida evaluations, and the widely circulated quality standards of the Joint Committee on Standards (1994) are also relevant. According to the Joint Committee, quality in evaluation can be assessed in relation to four interrelated criteria: accuracy, feasibility, propriety and utility. While the first concerns factual correctness and adequacy of the information provided by an evaluation, feasibility and propriety refer to the practicality of the evaluation and its conformity to ethical standards respectively. Finally, utility refers to the usefulness of an evaluation in relation to the problem it is intended to solve (cf. Sida 2007, p 24).

In this study we are mainly concerned with quality in relation to the criteria of accuracy and utility. More precisely, we focus on the following issues:

1. the quality of the TOR and the evaluation questions, and the extent to which the evaluations respond to them;
2. the quality of the evaluation research designs, including methods for data collection;
3. the quality of the results information and the analyses of implementation processes provided by the evaluations; and
4. the quality of the conclusions, recommendations and lessons learned that are contained in the reports.

We proceed on the assumption that the same quality standards can be applied to all evaluations, regardless of purpose and context. This assumption can be questioned. There is a strong case to be made for applying quality standards selectively. If, for example, an evaluation is primarily commissioned to document experiences for organisational learning, the attributes that make it easily readable and understandable might be of great importance. If, however, an evaluation is commissioned to assess results before a decision is made on whether to continue a programme, the intended readers may be few and hence the communicative aspects less important. On the other hand, quality standards referring to methodological choice, data and results, and the drawing of conclusions are always important regardless of context and purpose.

The model in Box 1 sets out a general framework for assessing evaluation quality in relation to the four issues above. On the basis of this model we identified no less than 64 separate aspects or elements that we considered relevant to our task. Annex 1 contains our assessment format with questions relating to each of these 64 elements. Of the questions, 17 refer to background characteristics, 7 to a description of the methodology and the remaining 40 to aspects of an evaluation that are directly relevant to an assessment of its quality.[2]

Box 1. Extended model to assess the product, process, and information request quality of evaluation reports

Descriptive category	Main issues assessed/described
Description of system aspects of the evaluation	• Cost of the evaluation • Sector, nature of evaluated object • Region • Evaluators/evaluation team • Host country participation
Description of methodology	• Basic evaluation question(s) • Evaluation design • Evaluation methods • Use of data collection instruments
Assessment of methodological choices	• TOR and basic question(s) • Design and methods • Validity and reliability • Methodological choices • Data collection instruments
Assessment of evaluative findings	Reliability of assessment of management and implementation Reliability of assessment of outputs, outcomes and impacts

[2] The analytical framework adopted in this study is similar to that used by Forss and Carlsson 1997 and Forss and Uhrwing 2003.

Descriptive category	Main issues assessed/described
Assessment of conclusions and recommendations	Conclusions that are based on evidence
	Recommendations that follow from value premises, data analysis and conclusions
	Lessons learned that are clear and succinct and follow from empirical observations

Each of the reports was assessed against the 40 quality criteria, and the assessment of each one was summarised as a rating on a six-point scale ranging from 'excellent' to 'very poor'. The aggregation of ratings that refer to different quality criteria into a combined overall quality rating was avoided, as a good rating according to one criterion, such as clarity of presentation, does not necessarily compensate for a poor rating by another criterion, such as analysis of attribution. Although, to some extent, strengths seem to go hand in hand with strengths and weaknesses with weaknesses, it was not considered practically useful to construct a composite quality index.

An Excel master sheet was developed in which each evaluation report was given a row and each quality indicator a column. As all the ratings were plotted on this sheet, it became our main database for this study (see Annex 2). In the course of reading and discussion, the team members also took note of examples of "good practice" and other instructive solutions to evaluation problems. Examples of these are presented in text boxes throughout the report.

Each of the reports was carefully read and rated by at least two of the team members. The first reading was carried out individually. We then met and compared our assessments in order to agree on a consolidated opinion. There were initial differences of opinion in many cases, but, through discussion, we were usually able to arrive at a common understanding and joint conclusions. On the whole, we believe that the assessments presented in this study are accurate and fair.

This is not to say that our assessments are beyond dispute. The fact that all the members of our team are experts in evaluation rather than experts in the various substantive fields discussed in the evaluations is obviously a potential source of bias in itself. It is quite possible that experts in those fields would assess the strengths and weaknesses of the reports differently.

There is also a risk that we have put too much emphasis on bureaucratic neatness and academic accuracy, forgetting at times that evaluation is primarily a practical decision-making tool. As it turns out, assessing the quality of evaluation reports is not the same as producing such reports. Furthermore, our individual understanding of the reports tended to change as we discussed them, and it might have continued to do so had we allowed the discussion to go on. The negotiated consensus that we present in this report is not necessarily the last word on the quality of those reports. Our assessments should be taken as a contribution to a debate that can, and should, continue.

1.4 Quality Questions

From the four major interrelated criteria described above (cf. 1.3.), we developed eight questions to discuss the quality of the sample evaluations. As Sida's evaluation system has been in place for many years it seems reasonable to expect that most evaluations would pass a quality test. It should also be expected, for a variety of reasons, that some would fail. What percentage of Sida's evaluations can be rated as "satisfactory" in respect of the different quality criteria? The rating uses a six-point scale, with satisfactory being a rating in one of the upper three categories.

Question 1. Do Sida evaluations adequately address the evaluation questions formulated by Sida in the TOR?

Evaluations are commissioned for a purpose, which is supposed to be clearly spelled out in the TOR. A number of questions follow from the purpose, based on the five main evaluation criteria – effectiveness, efficiency, impact, relevance and sustainability – that the evaluation is meant to answer. Not all TOR require an assessment of all five criteria and the evaluator is supposed to discuss the evaluation questions before developing a methodology to answer them. While much could be said about the importance of well-written TOR, this question focuses on the extent to which the evaluation reports answer the questions posed in the TOR.

Question 2. Do Sida evaluations provide valid and reliable information on efficiency, effectiveness, impact, relevance and sustainability?

According to the OECD/DAC Evaluation Quality Standards, evaluation is defined as a study of efficiency, effectiveness, impact, sustainability and relevance (OECD/DAC 2007). Hence, as these reports are entitled "evaluations", they must, by definition, contain information in these areas. As explained in the Sida Evaluation Manual, not all five criteria need to be covered in every evaluation: "the policy requirement is rather that none of them should be put aside without a prior assessment of their relevance" (Sida 2007, p. 28). However, if the evaluation system as a whole is expected to provide sufficient information on the five dimensions mentioned above, the dimensions need to be applied frequently and evaluations should contribute valid and reliable findings.

Question 3. Do Sida evaluations contain a clear and consistent analysis of attribution and explain how and why the interventions contributed to the results?

Question 2 addressed the analysis of results, and in practice this should include an analysis of how the changes are brought about. This is not always the case and methods for drawing conclusions on issues such as effectiveness

and impact can vary a great deal. In order for an evaluation to be useful, presentations of reliable results should, as far as technically possible and practically feasible, be accompanied by an analysis of how the change was brought about. We have therefore introduced this question, which focuses on an analysis of how the intervention contributed to the results (in terms of, for example, impact or outcome).

Question 4. Do Sida evaluations have an appropriate design for impact evaluation?

Evaluations can take many different forms: sometimes it is possible to design experimental studies with randomised test groups and control groups and at other times case study designs or narrative analysis are more suitable and respond best to the TOR[3]. Evaluators choose from interviews, surveys, observations and document analyses as their main data collection methods. As the subjects under evaluation are so different we should expect a variety of approaches to the evaluation task.

Question 5. Is the evaluation process in Sida evaluations well documented and transparent so that readers can make an independent assessment of validity and reliability?

Evaluation is also defined as systematic inquiry, which means that the methods of the social sciences should be used. An evaluation is often more useful if the process is transparent, making the process of inquiry visible to the readers. Many evaluations, however, try to be short and concise, and the readers might be more interested in the conclusions than the methods. Even so, it seems reasonable to expect that most evaluation reports inform their readers of what they have done and why their findings should be trusted.

Question 6. Do Sida evaluations include a valid and reliable analysis of the management of interventions?

Evaluations are expected to lend support to the decision-making process, for example, by suggesting how the management of interventions could be improved. Even if the focus is on the results, it is important to analyse how the results were produced, rather than to treat the implementation process as a black box. The TOR often expect evaluators to document the implementation and to suggest reforms of organisational structures and processes. We would therefore expect most of the evaluations to include a careful analysis of the implementation so that they can make recommendations for the future as well as promote learning.

[3] A study by World Bank evaluation personnel analysed how evaluation design can vary in the development context: Bamberger et al (2004).

Question 7. Do Sida evaluations provide clear and focused recommendations for specified target groups?

In many cases an evaluation is intended to support decisions. This means that an evaluation should identify and recommend a course of action. Many guides have been written on how to develop useful recommendations (for example Patton 1997). An important aspect is to identify the various stakeholders and suggest recommendations that are within their mandate and scope for action.

Question 8. Do Sida evaluations document interesting and useful lessons learned from the interventions that were evaluated?

One of the two main purposes of evaluation is to contribute to learning: within Sida, among partners, and among people interested in development cooperation. Lessons learned are "generalisations based on evaluation experiences" (Sida 2007, p. 110) and "general conclusions with a potential for wider application and use" (Sida 2007, p. 87). The degree of generalisation may vary from case to case, however, and it may not be possible for all evaluations to formulate new lessons for a wider community of development practitioners.

2 The Evaluation Sample

2.1 Introduction

This chapter presents the sample of 34 evaluations reviewed in this study. It answers the following questions:

- How does Sida's evaluation system work?
- How was the sample chosen?
- What is being evaluated?
- When are the evaluations carried out?
- How much do the evaluations cost?

2.2 Sida's Evaluation System

Sida evaluations are commissioned by the thematic and regional departments and the Swedish Embassies in partner countries, as well as by Sida's Department for Evaluation (UTV). Each department and embassy conducts evaluations within its own area of responsibility. UTV, which is an independent function reporting directly to Sida's Director General[4], conducts strategic evaluations of wider scope, and also advises the thematic and regional departments on their evaluation work.

As a basis for its advisory services, every year UTV assembles the evaluation plans of Sida's departments and the Swedish embassies in partner countries into an overall annual Sida evaluation plan. In recent years, this plan has included approximately 40 evaluations.[5] As they are completed, the evaluations figuring in the plan are published in the Sida Evaluations series (SE). All the evaluations published in this series can be ordered directly from Sida or downloaded from Sida's website (www.sida.se).

While satisfying Sida's definition of the concept of evaluation[6], some of the items in the SE series are fairly light-weight types of studies that would, in some other organisations, have been regarded as 'reviews' or even as moni-

[4] Since February 1, 2008 UTV reports to Sida's Director General. Prior to that it reported to Sida's Board of Directors, a body that no longer exits.

[5] As Sida's line departments and the Swedish embassies in partner countries sometimes fail to report their evaluations to UTV, the number of evaluations conducted by Sida each year is probably somewhat larger than the number of evaluations recorded in Sida's annual evaluation plan.

[6] Sida defines the concept of evaluation as follows: "…an evaluation is a careful and systematic retrospective assessment of the design, implementation, and results of development activities." Looking Back, Moving Forward. Sida Evaluation Manual, 2007, p. 11.

toring reports rather than as genuine evaluations. For reasons of transparency, however, Sida interprets the concept of evaluation generously and usually prefers to publish than not to publish. UTV would normally not object if a department wants a particular evaluation study to be published as a *Sida Evaluation*. The responsibility for maintaining the quality of the series rests with all the departments contributing to it rather than with UTV alone, although UTV has the authority to say no.

Note also that the SE series does not include evaluations that Sida conducts jointly with other donors. SE consists of studies initiated by Sida alone and most of the evaluations in the series are project evaluations rather than evaluations of programme support. Recommendations are often directed at Sida's cooperation partners in the host country government, but it is not clear to what extent this advice has been explicitly requested. Presumably it is used by Sida staff as a basis for dialogue with their host country counterparts. Less than half of the evaluations reviewed in this study had some form of participation from the host country in the evaluation team.

2.3 The Sampling Process

This study is based on an analysis of a sample of SE reports. As we wanted an assessment of current evaluation quality, we decided to define our sampling universe as the SE reports published during 2003, 2004 and 2005. This came to a total of 96 reports in Sida's evaluation database.

From this population we selected 34, which was just over 30% of the total. The decision to restrict the sample size in this way was mainly practical: a sample of 30% or more could be expected to be representative of the total population, while less than 30% might be questioned as atypical. As a quality assessment of this kind involves a lot of work we did not want to deal with more evaluations than required for convincing conclusions.

The selection of the 34 reports was a process in several steps. As it was necessary to try out the assessment model, five reports were selected as pilots. In order to prepare the ground for a planned, later study of country-specific ways of using M&E in Mozambique and Vietnam, four of the pilots were evaluations referring to these countries. Of the remaining 29 reports, 24 were chosen at random with the help of a table of random numbers and 5 were chosen because they referred to Mozambique and Vietnam. Thus, in the total sample of 34 there were no less than 9 evaluations dealing with Mozambique and Vietnam.

Furthermore, while the study was well under way, we decided to take out four UTV evaluations that were part of the original sample and replace them with four evaluations from the line departments, also chosen at random. We did this because we felt that comparing the often relatively light-weight and low-cost evaluations from the line departments with the more ambitious

UTV evaluations was not quite fair. The four evaluations from UTV were used for illustration but were not rated along with the others. The evaluations included in the rating exercise are all listed in Table 1.

Table 1. Evaluation reports that were assessed in the review

Evaluations assessed in the pilot phase (n=5)	
SE 02/12	Strengthening the Capacity of the Office of the Vietnam National Assembly
SE 02/35	Implementation of the 1999–2003 Country Strategy for Swedish Development Cooperation with Vietnam
SE 03/35	Sida Support to the University Eduardo Mondlane, Mozambique
SE 04/14	Sida's Work Related to Sexual and Reproductive Health and Rights 1994–2003
SE 04/29	Mozambique State Financial Management Project
Evaluations assessed in the main phase (n=29)	
Evaluations from Mozambique and Vietnam (n=5)	
SE 02/06	Research Cooperation between Vietnam and Sweden
SE 02/07	Sida Environmental Fund in Vietnam 1999–2001
SE 03/09:1	Contract-Financed Technical Cooperation and Local Ownership: Botswana and Mozambique Country Study Report
SE 03/29	Institutional Development Programme (RCI) at the Ministry of Education in Mozambique
SE 04/35	Local Radio Project in Vietnam 2000–2003
Evaluations chosen at random (n=24)	
SE 03/01	Sida Support to PRONI Institute of Social Education Projects in the Balkans
SE 03/05	Zimbabwe National Network of People Living with HIV/AIDS
SE 03/11	Development Cooperation between Sweden and the Baltic States in the Field of Prison and Probation
SE 03/12	Three Decades of Swedish Support to the Tanzanian Forestry Sector: Evaluation of the Period 1969–2002
SE 03/19	Sida's Health Support to Angola 2000–2002
SE 03/25	Aid Finance for Nine Power Supervision and Control Systems Projects, an Evaluation of SCADA Projects in Nine Countries
SE 03/27	Africa Groups of Sweden's Programme in Malanje Province – Angola 1999–2002
SE 03/38	The Swedish Helsinki Committee Programme in the Western Balkans 1999–2003
SE 03/41	Sida funded Projects through UNICEF-Bolivia, 1989–2002
SE 04/04	Management Audit of the Swedish Red Cross
SE 04/10	Zimbabwe Aids Network
SE 04/18	The Regional Training Programme in Design, Installation, Administration and Maintenance of Network Systems (DIAMN)
SE 04/21	Water Education in African Cities United Nations Human Settlements Program
SE 04/22	Regional Programme for Environmental and Health Research Centres in Central America
SE 04/23	Performing Arts under Siege

Evaluations chosen at random (n=24)	
SE 04/24	National Water Supply and Environmental Health Programme in Laos
SE 04/32	Environmental Remediation at Paddock Tailing Area, Gracanica, Kosovo
SE 04/33	Swedish Support to Decentralisation Reform in Rwanda
SE 04/38	Sida's Work with Culture and Media
SE 04/36	Life and Peace Institute's Projects in Somalia and the Democratic Republic of Congo
SE 05/04	Regional Training Programme in Environmental Journalism and Communication in the Eastern African Region
SE 05/14	What Difference Has It Made? Review of the Development Cooperation Programme between the South African Police Service and the Swedish National Police Board
SE 05/13	Integrating Natural Resource Management Capacity in South East Asia
SE 05/16	Partnership Evaluation of Forum Syd 2001–2003

2.4 The Evaluated Interventions

The reader will have noticed that we write about the "intervention" or "object" that is evaluated. These are blanket terms covering policies, programmes, projects, core funding of organisations, etc. Of the 34 sample reports, 19 deal with projects, 8 are programme evaluations, and the remaining 7 are policy evaluations and organisational assessments. Note that the distinction between programmes and projects is not always clear. SE 04/29, for example, which deals with the Mozambique State Financial Management Project, does not appear to have a different kind of object to SE 03/29, which according to its title is an evaluation of an institutional development programme in the same country. As the terms are used by the evaluations in the sample, projects and programmes are often much alike in terms of objectives, time frame, implementation and budget consequences.

As Sida, together with most other bilateral development cooperation agencies, is moving away from project financing to wider forms of cooperation such as sector support and general budget support, one might have expected to find more evaluations of such forms of cooperation in the sample. As explained above, however, evaluations of general budget support and the like are usually joint evaluations that are not published in the SE series. Furthermore, although there has been a change towards sector support and general budget support, Sida funds are still allocated to projects and project-like programmes for the most part.

2.5 Timing of the Evaluations

The assessment model includes a question about the timing of the evaluation in relation to the evaluated object. The key distinction is that between evaluations of ongoing interventions and evaluations of completed interventions.

It was not always easy to classify the sample evaluations in relation to this distinction however. SE 03/12, which deals with 30 years of Sida support to the forestry sector in Tanzania, is one example. Many of the projects supported by Sida had come to an end long before the evaluation, others had been completed only recently, and still others were ongoing. As a whole, the evaluation fell into both categories.

Nevertheless, relatively few sample evaluations were carried out after the intervention had come to an end. The activities under review were usually ongoing. This is worth noticing as it means that outcomes, impacts and sustainability could not be properly assessed. Assessments of those types of results can only be made when the intervention has existed for some time or after it has come to an end. However, most of the sample evaluations had been conducted too early for an accurate assessment of such results to be possible. Questions about the likelihood of intended and unintended future impacts and long-term sustainability can and should, of course, be raised in early evaluations, but an assessment of the likelihood that something will happen in the future is not the same thing as an evaluation seeking to find out if outcomes and impacts have actually occurred as expected.

The question of the timing of the evaluations would also seem to be relevant to an assessment of the quality of the overall evaluation system. There are good reasons to undertake evaluations during the implementation of a programme in order to provide information for management. However, in order to promote learning regarding factors that are likely to affect long-term results it is also necessary for evaluations of completed interventions to be undertaken. According to our findings there is a lack of such evaluations in Sida's evaluation portfolio. Assuming, as we usually do, that information about the results of past efforts can help improve current initiatives, this would seem to be a significant weakness of the evaluation system as a whole.

2.6 Resources Spent on Evaluations

The average budget for the evaluations in this assessment was 780,000 SEK, with individual evaluations costing between 116,000 SEK and 2,642,000 SEK. The costs included consultants' fees as well as travel costs and accommodation for meetings and field trips. While the budget for some evaluations seemed appropriate, others had budgets that severely limited the amount of time that could be spent in the field.

Table 2. Evaluation costs

The five most expensive evaluations in the sample	SEK
SE 04/29 Mozambique State Financial Management Project	2,642,000
SE 04/38 Sida's Work with Culture and Media	1,492,000
SE 04/14 Sida's Work Related to Sexual and Reproductive Health and Rights 1994–2003	1,160,000
SE 04/36 Life and Peace Institute's Projects in Somalia and the Democratic Republic of Congo	1,093,000
SE 03/35 Sida Support to the University Eduardo Mondlane, Mozambique	1,054,000
The five least expensive evaluations in the sample	
SE 03/05 Zimbabwe National Network of People Living with HIV/AIDS	116,000
SE 04/10 Zimbabwe Aids Network	122,000
SE 04/22 Regional Programme for Environmental and Health Research Centres in Central America	161,000
SE 04/24 National Water Supply and Environmental Health Programme in Laos	161,000
SE 05/04 Regional Training Programme in Environmental Journalism and Communication in the Eastern African Region	199,000
Average cost of evaluations in the sample	780,000

Source: SE fact sheets and supplementary information from Sida

There is not always a clear connection between budget and time and the expectations expressed in the TOR. Different evaluations pose different challenges and make different demands, for example, sometimes focusing primarily on project management, and at other times involving analyses of factors enabling or preventing poverty reduction at societal levels. It is necessary for evaluators to assess the time available and spend it as productively as possible on a range of different tasks: choice of methodology, data collection through meetings with key informants and field work, data analysis, report writing and so on. It would seem likely that time and budget would have an impact on the quality of the evaluation, and it is therefore interesting to note that we did not find a clear and consistent correlation between budget and quality in this assessment.

As we take a closer look at the relationship between quality and costs, however, the lack of such a correlation is not surprising. As already suggested, the critical question is whether the resources invested in the evaluation are sufficient to produce a study that satisfies the requirements set down in the TOR. The total amount of money invested in the study tells us nothing about the quality of the study. As a buyer of evaluation services, Sida must try to make sure that the TOR are realistic given the resources that can be invested in the evaluation and that the resources are adequate given the TOR. In evaluation, as elsewhere, ensuring quality means mutually adjusting means and ends.

3 Questions and Answers

3.1 Introduction

The previous chapter described the nature of the evaluations in the sample. This chapter analyses the information presented in the evaluations and assesses to what extent it matches the TOR:

- Do the evaluations provide relevant and adequate answers to the questions in the TOR?
- What types of results information do the reports contain?
- Do they provide accurate presentations of what happened during implementation?

Sida's evaluation manual, Looking Back, Moving Forwards (2007), refers to five well-established evaluation criteria: relevance, effectiveness, efficiency, impact and sustainability (Box 2, below). The first part of this chapter will discuss these criteria. There are also a number of common evaluation issues that focus on various aspects of planning, implementation and the results of interventions, and these will be discussed in the latter part of the chapter. It should be emphasised that not all evaluations need to discuss achievement of all the evaluation criteria and address as many questions as possible. We are not arguing that the best evaluation report is the one that answers as many questions as possible. An assessment focusing exclusively on impact could produce an excellent report. The same is true for an evaluation of management capacity, organisational systems, cost-effectiveness or long-term sustainability. It is the TOR that should decide the scope of an evaluation. A good evaluation should answer questions raised in the TOR.

The focus and perspective of an evaluation is also likely to be determined by the overall purpose of the study as understood by the evaluators through interaction with stakeholders. If the overall purpose of the evaluation is accountability – providing feedback to principals on the value of the investment – the focus will in many cases be on measuring and documenting short- and long-term results. The donor may often be less interested in how well a project was planned, organised and implemented and more concerned with what was achieved through the intervention. If the overall purpose of an evaluation is organisational learning, its focus will be different. It will in many cases be more participatory and focus more on implementation processes – trying to understand what factors facilitate and constrain performance.

3.2 Terms of Reference – The Starting Point

We found that most of the evaluations in our sample addressed the questions raised in the TOR, although not necessarily providing satisfactory answers (see Table 3 below). Only six were less than adequate in terms of coverage and none was deemed to have significant shortcomings. Evaluation teams always present draft reports to Sida, and the programme officer, alone or in consultation with other stakeholders, assesses whether the evaluators have responded to the TOR. If they have failed, they are to be told so in no uncertain terms. Hence it is not surprising that the end product corresponds fairly well to the TOR.

Table 3. Assessment of response to terms of reference

	1	2	3	4	5	6	N/A	Total
Does the evaluation respond to the questions in the TOR?	0	2	4	13	9	6	0	34

Key to ratings: 1 – very poor (or not done at all), 2 – significant problems, 3 – not quite adequate, 4 – minimally adequate, 5 – adequate, 6 – excellent, NA – not applicable, the question was irrelevant to that evaluation, or the issue could not be assessed because of a lack of information.

Source: The authors' assessment of 34 evaluation reports

The TOR were not always clearly formulated and well focused. In many cases they asked for more than the evaluators could possibly deliver, given the time and resources available to them. Our overall assessment of the TOR for the evaluations examined in this study is that they were not very good. No report had TOR that we rated as "excellent" and fewer than half of them were considered "adequate". One in five was deemed more or less inadequate.

Many TOR failed to describe the overall purpose of the evaluation – its intended use – clearly. Instead of providing the reader with an explanation of the rationale for the study they proceeded directly to the evaluation questions, which in many cases were not only quite detailed but also numerous. A problem with TOR designed in this way is that they make it difficult for the evaluators to adapt to unexpected findings or factors during the research process. TOR that prescribe a particular methodology can be problematic in the same way, since they may prevent evaluators from flexibly exercising their own best judgement, encouraging them instead to mechanically adapt to the client's expectations, regardless of the results.

Most of the TOR presented the evaluators with a broad range of standard questions about impact, effectiveness, relevance, sustainability, etc. Such questions are usually demanding and difficult to answer with a reasonable degree of precision, especially with limited resources and in a short period of

time. It seems, however, that a majority of the evaluation teams adopted Sida's TOR without any discussion of relevance, feasibility, the need for a clearer focus or a concentration of resources. It was not common for evaluation teams to present an independent interpretation of the TOR in the report at any rate. The evaluation questions formulated in the introductory chapters of the reports were most often copied directly from the TOR, with only slight changes of wording. Only in a few reports were they further interpreted, operationalised, or assessed with regard to their relative importance to the evaluation purpose. Reinterpretations of the evaluation questions through an explicit analytical model or conceptual framework were very much the exception.

One therefore does not get the impression from reading the sample reports that the TOR were closely discussed by the Sida programme officer and the consultants at the beginning of the evaluation process. In an evaluation of the implementation of the Swedish country strategy for Vietnam (SE 02/35) the evaluators sought clarification of the TOR from Sida on a number of points, but this is the sole example of its kind.

Table 4. Assessment of the evaluation question(s)

	1	2	3	4	5	6	N/A	Total
Are the TOR clear and focused?	0	1	5	12	16	0	0	34
Does the evaluation interpret and focus the task as defined in the TOR?	10	3	5	6	8	2	0	34
Is the basic question clearly stated in a specific section?	7	4	5	5	1	2	10	34
Can the informed reader arrive at an understanding of the basic question?	0	1	2	13	14	4	0	34

Key to ratings: 1 – very poor (or not done at all), 2 – significant problems, 3 – not quite adequate, 4 – minimally adequate, 5 – adequate, 6 – excellent, NA – not applicable, the question was irrelevant to that evaluation, or the issue could not be assessed because of a lack of information.

Source: The authors' assessment of 34 evaluation reports

Some agencies (the EC for example) request that an inception report be prepared as a first step in an evaluation. In this report the evaluators are expected to give their interpretation of the evaluation questions in the TOR and present their choice of evaluation design and data collection methods. This is not a mandatory requirement for Sida evaluations but the inception report procedure was used in a few of our cases. The TOR for SE 04/36 contains the following requirement.

> "The Selected Consultant is asked to begin the assignment by preparing an inception report elaborating on the feasibility of the scope of the evaluation, the methodology for data collection and analysis, the detailed and operational evaluation work plan (including feedback workshops). During this stage it is important that information is sought from the Institute's offices in Nairobi and Bukavu and not only from the office in Uppsala." (SE 04/36: Life and Peace Institute's Projects in Somalia and the Democratic Republic of Congo, Annex 1: TOR.)

Such investments in early clarification of the evaluation questions often pay off later. In small evaluations with few and straightforward questions, an inception report might introduce an unnecessary loop – adding time and costs but not much value. In complex evaluations with a broad range of difficult questions, however, an inception report is often a useful tool to facilitate communication about the focus of the assignment and about how realistic or evaluable the questions are.

An inception report allows the evaluator to make an informed up-front judgement of the feasibility of the assignment. In most cases such a report will be an integral part of the contract. If an inception report is required, the TOR can often be relatively brief, focusing on issues that need to be settled before conducting the evaluation. If an inception report is not required the TOR would normally be more detailed.

A majority of the TOR in this study state that the evaluation report should not exceed a limited number of pages. Such a requirement is common even when the evaluation questions are numerous and complex. Limiting the size of the report in advance of the evaluation process seems not only unnecessary but also potentially harmful to the quality of the results. It is notable, however, that while some evaluators comply with this requirement, others disregard it completely.

3.3 Results Assessments

We will now look at how the evaluations in our sample deal with the five OECD/DAC evaluation criteria: relevance, effectiveness, efficiency, impact and sustainability. Two questions are addressed: 1) to what extent are the five evaluation criteria covered by the sample evaluations (and their TOR)? 2) What is the quality of the assessments? Box 2 below provides compact definitions of the criteria.

Box 2. Five evaluation criteria – the basic questions evaluations are expected to answer

Evaluation criterion	Specification
Efficiency	The extent to which the costs of a development intervention can be justified by its results, taking alternatives into account
Effectiveness	The extent to which a development intervention has achieved its objectives, taking their relative importance into account
Impact	The totality of the effects of a development intervention, positive and negative, intended and unintended
Relevance	The extent to which a development intervention conforms to the needs and priorities of target groups and the policies of recipient countries and donors
Sustainability	The continuation or longevity of benefits from a development intervention after the cessation of development assistance

Source: Looking Back, Moving Forward. Sida Evaluation Manual (p. 25)

As explained carefully in Sida's evaluation manual, each of these criteria can be applied to every development intervention and each one represents an important results dimension that needs to be considered before it can be decided whether or to what extent an intervention should be regarded as a success. It is not Sida policy, however, that all evaluations must cover all the criteria. There are situations in which it is right to ignore one or several criteria, or so it is argued. In other words, the existence of evaluations that do not apply all five criteria is not, in itself, a quality problem. On the contrary, it could be seen as a strength that some evaluations focus on just one or two, but do it well.

As shown in Table 5, however, the majority of the evaluations in our sample do in fact refer to all five criteria, though in many cases only superficially. This reflects the fact that most of the TOR provide a comprehensive mandate for the evaluation, without much discrimination between the criteria. The most commonly covered criteria was that of effectiveness followed by impact, relevance, sustainability and efficiency, in that order.

This inclusive approach is probably due to UTV's efforts to popularize the OECD/DAC model over a period of several years. Sida's evaluation policy states that the relevance of all five criteria should be considered every time an evaluation is planned and Sida's evaluation manual provides guidance for how this can be done. We can assume that every Sida programme officer who is charged with the task of writing TOR for an evaluation is familiar with the five criteria, and it is likely that most adopt all five as an easy solution to what could otherwise become a rather difficult selection problem. Whatever the explanation, however, the wholesale adoption of the OECD/DAC model ensures a broad analysis in the reports, which in itself represents strength, though it may lead to a lack of focus and prioritisation and have a negative effect on the quality of individual evaluations.

Table 5. Coverage by evaluation criteria

Evaluation criteria	No. of reports with applications	No. of reports without applications
Efficiency	29	5
Effectiveness	31	3
Impact	30	4
Sustainability	29	5
Relevance	30	4

Source: Assessment of the sample evaluation reports

Compared to Table 5, which merely registers whether the criteria were discussed or mentioned at all, Table 6 presents a summary of our assessments of how well results were analysed in relation to the criteria. Before turning to a review of the criteria in turn, a few overall comments are required:

- The assessment of the relevance of interventions was generally found to be more accurate and adequate than the assessments referring to the other criteria, although it usually only covered certain aspects of what we mean by relevance (cf. 3.3.1). This is encouraging inasmuch as it means that the evaluated interventions were assessed from a broader development perspective and analysed from the perspectives of key stakeholders. On the other hand, an assessment of relevance is rarely good enough by itself. An analysis of actual or potential effects would usually also be required.
- Intervention effectiveness is considered in 31 of the 34 reports, and impact in 30. Many of the evaluations that draw conclusions regarding intervention, effectiveness did not give the issue of attribution sufficient consideration however, i.e. they did not provide sufficient evidence that the documented changes were due to the evaluated *intervention*. As both effectiveness and impact refer to the extent to which interventions have actually made a difference, the lack of attention to the attribution issue is rather surprising.
- Information regarding the efficiency of the evaluated interventions was deemed "less than adequate" in all but 8 of the reports – either because the analysis was weak or because it was missing altogether (although seemingly relevant). The assessments of efficiency were also found to be less accurate generally than the assessments referring to the other criteria.

Table 6. Results assessments in evaluation reports

	1	2	3	4	5	6	N/A	Total
Is there an accurate assessment of efficiency?	7	5	10	4	2	1	5	34
Is there an accurate assessment of effectiveness?	2	4	3	9	12	0	3	34
Is there an accurate assessment of impact?	6	5	3	9	4	3	4	34
Is there an accurate assessment of sustainability?	4	5	4	11	4	1	5	34
Is there an accurate assessment of relevance?	4	2	2	8	11	3	4	136

Key to ratings: 1 – very poor (or not done at all), 2 – significant problems, 3 – not quite adequate, 4 – minimally adequate, 5 – adequate, 6 – excellent, NA – not applicable, the question was irrelevant to that evaluation, or the issue could not be assessed because of a lack of information.

Source: The authors' assessment of 34 evaluation reports

3.3.1 Relevance

Relevance refers to the extent to which intervention objectives and activities are in line with the needs and priorities of target groups and with the policies of recipient countries and donors. The two latter aspects can in many cases be addressed through a straightforward analysis of easily accessible documents (comparing programme documents with national plans of the recipient country and Swedish policies, respectively), although it is of course always important to consider the degree to which the documents are actually taken seriously by their sponsors. Assessing the interventions in relation to the priorities and needs of target groups, however, is usually a much more complex task. Not surprisingly, most assessments of relevance focused on the official documents. Questions concerning the degree of consistency of the intervention with target group interests were rarely addressed.

This bias towards the documented views of governments and donors is not reflected in our ratings. We did *not* give a lower quality rating to reports that failed to discuss the potential usefulness of the intervention from the point of view of target groups than to those (very few) that provided such an analysis. Other differences were felt to be more important. For example, while some of the evaluations limited themselves to a fairly narrow analysis of consistency with officially proclaimed donor and country goals and objectives, others ventured into a more complex and, in our estimation, rather more useful discussion of the intervention in relation to its urgency in relation to needs, and its value in relation to alternative and potentially more appropriate uses of the same resources.

The nature of the data was important to our assessments in more than one way. In SE 05/16, for example, an evaluation of interventions sponsored by Forum Syd, relevance was analysed in terms of:

(1) beneficiaries' needs

(2) the partner civil society organisation's (CSO) goals

(3) Forum Syd objectives

(4) Sida objectives

The analysis thus referred to all the major stakeholders and to needs as well as to objectives. However, the information about the relevance of the reviewed interventions to the affected target groups was provided by local and Swedish partner organisations rather than from these groups themselves. Information gathered in this way can of course not be taken at face value, but should be understood for what it is, namely interested and possibly partial and biased statements by one stakeholder group with regard to another.

3.3.2 Efficiency

For a donor like Sida, questions about efficiency – broadly speaking value for money – are almost always likely to be of interest and relevance and, not surprisingly, most of the TOR in our sample included such questions. In most of the reports, however, the assessment of efficiency was technically quite weak. While all the reports included information about the resources spent on the intervention, very few provided a systematic assessment of the value of the benefits (outputs, outcomes, impacts) of the evaluated intervention in relation to the costs of producing them.

The fact that questions about efficiency are technically demanding is probably one of the main reasons for the lack of competent efficiency assessments in the sample reports. Where assessments of efficiency are made they tend to focus on questions about productivity or internal efficiency (Vedung 1998: 254 ff.). Assessments of costs in relation to outcomes or impacts, which tend to be more complex, are less common. The standard critical observations about efficiency concern such things as excessive administrative expenditure or the need to reduce unit costs.

An evaluation of an initiative to integrate natural resource management capacity in South East Asia (05/13) is a good example. The sections of the report that deal with efficiency are all about administrative overheads and the possibility of reducing costs. Similarly, in an evaluation of a regional training programme in Sri Lanka (04/18), the assessment of efficiency concerns unit costs (cost per student, including travel costs). Questions about the extent to which more and better development outcomes or impacts might be achieved by alternative uses of the available resources are rarely discussed in the sample reports.

The evaluation of the cooperation programme between the South African Police Service and the Swedish National Police Board (SE 05/14) is one of the evaluations in which costs are assessed in relation to outcomes as well as outputs. The report concludes as follows:

> "As to the analysis of cost in relation to outputs and outcomes as revealed by the accounts for the Swedish contribution and the detailed scrutiny of each project, the results yielded must on the whole be said to give good value for money." (p. 7)

As the evidence behind this statement is not given in the report, the reader cannot assess the validity of the assessment. The case is not unique. All too often, conclusions like the one above are presented without supporting data.

3.3.3 Sustainability

Few evaluations apply the sustainability criterion well, and five reports do not discuss sustainability at all. Although sustainability – what will happen with the intervention or its benefits when the external assistance comes to an end – tends to be regarded as an important issue in most of the evaluated projects and programmes, the sample TOR do not always include or prioritise its analysis.

We should keep in mind that unlike assessments of relevance, efficiency, impact and effectiveness, assessments of sustainability are projections into the future. In most cases the issue of sustainability is analysed in hypothetical terms – A is likely to be sustained provided that B remains in place and C does not happen, etc. The analysis draws on general experience about what sustainability seems to require with regard to things like stakeholder participation, the role of government or civil society structures in implementation, the ability of partner organisations to cover recurrent costs, etc. It is a common point that the chances of structures or benefits being sustained into the future are likely to increase if the right structures of local ownership and management are built or put in place early on in the intervention process. The following statement from an evaluation of Sida-funded projects with UNICEF in Bolivia (03/41 p 35) is typical: "The overall conclusion of the evaluation team is that the greatest likelihood of sustainability is found in the projects that have become integrated with national policies and programmes."

Box 3. Examples: Analysis of sustainability

The evaluation of Africa Groups of Sweden's Programme in Malanje Province – Angola provides a brief analysis of sustainability:

> "One of the most important areas regarding sustainability is the question of whether the social organization promoted in the programme will have enduring effects. The consultants were not able to prove that the interest groups formed in the temporary settlements survived when the IDPs returned to their origins. However, the evaluation team found some evidence that the community organizations in some cases had survived. The fact that Malanje Antena is comprised of local individuals is some warranty for sustainability."

A similar conclusion is found in an evaluation of support to the Office of the National Assembly in Vietnam (p. 42):

> "A strength of the project is the close relation to the operative work... The ideas and solutions provided through the project have, when found suitable, been integrated into the regular operations. In the field of public information several changes of this kind have taken place… It is more difficult to assess sustainability in other areas…. It is impossible at this stage of the cooperation to foresee what kind of future developments that may be attributed to this project."

Source: SE 03/27. Africa Groups of Sweden's Programme in Malanje Province – Angola 1999–2002 and SE 02/12. Strengthening the Capacity of the Office of the Vietnam National Assembly

Sustainability is a multi-layered concept with financial, technical, administrative and environmental dimensions. Few of the evaluations in the sample systematically cover the entire range of such types or dimensions. As suggested in Box 3, the assessments tend to be highly uncertain and tentative. Although an analysis of sustainability is, to some extent, inherently conjectural, a more systematic approach would in many cases have helped clarify the conclusions and make them less uncertain. With regard to sustainability, the sample reports seem to be based largely on subjective impressions and to consist of afterthoughts of analyses focusing on other criteria.

A report on research cooperation between Vietnam and Sweden (SE 02/06) provides an unexpected example of good practice. Although the term sustainability is not used and there is no separate discussion of the thing itself, two chapters that deal with capacity building for research and programme management respectively help us to understand key aspects of the sustainability issue. A clear argument is put forward regarding the extent to which the cooperation is contributing to capacity development and – by that route – sustainability. It is interesting that the analysis covers not only the focal organisations involved in the programme but looks at capacity building and sustainability in a wider context of national and regional research networks.

Box 4. Assessing sustainability: fragments of good practice

"Recognizing mutual interest ….. [I]t is rare that consultants, evaluators, or other experts make any fuss about [friendship]… Still, we all know how ubiquitous it is as a social force… Friendship can be a prime factor in processes of structural and normative change, and it appears to be one of the qualitative characteristics of cooperation on good programmes.

"The role of personalities – Cooperation is done by people, and it appears that some personal characteristics are more desirable… [I]t would seem appropriate to look for projects coordinators who possess … communication skills,… negotiation skills,… network building skills… and the ability to inspire trust and confidence among others…

"MOSTE programme ownership …. National ownership is a necessary precondition for an effective programme. … [T]he Vietnamese researchers [need to] possess a vision of the results they wish to achieve, have planned their cooperation, and keep track of progress….

"Phasing out strategy … It is useful to consider how and when a programme of cooperation should come to an end. Designing an exit strategy as part of a programme proposal can solve much anxieties, uncertainties and disappointment later on." (p. 34 ff.)

Source: SE 02/06. Research Cooperation between Vietnam and Sweden

3.3.4 Effectiveness and impact

Hardly any TOR in our evaluation sample do not include questions about effectiveness or impact. In recent years, most international donors, including Sida, have emphasised the need for more and better data about outcomes and impacts. There is a growing demand for well-documented impact assessments in order to prove to politicians and the public that development assistance makes a difference and that spending is well justified.

There are a number of formidable challenges regarding studies of impact and effectiveness. Effects in terms of, for example, poverty reduction are dependent on a number of factors, of which intervention is only one. Impact evaluations call for contextual knowledge and the analysis must take conditions and circumstances at many different levels into account. The results chain leading from the outputs of development intervention to its intended welfare effects can at times be quite long. Even when changes in outcomes can be measured it may not be possible to decide with much certainty whether they came about as a result of the intervention or if they were due to concurrent events. There are seldom quick and clear answers to impacts, as we will discuss in more detail below.

While measuring change can be a considerable problem in itself – baseline information, for example, is often lacking – the hardest questions tend to be those about causal attribution: Do the Sida-funded activities make a differ-

ence, and how can this be demonstrated? Would the situation have been the same without the Sida interventions? If there is a difference, how much of the outcome change can with reason be attributed to the Sida funding? Or, to put it differently, if there is a gross change in some outcome area, what is the net change in this area, i.e. the change produced by the intervention?

Few of the evaluations in the sample were able to provide precise and well-documented answers to questions about impact and effectiveness. As shown in Table 6 above, almost 50% of the reports were considered inadequate with regard to impact analysis, and among those that were considered adequate more than 50% were just barely adequate. The effectiveness ratings were better, but still not very positive.

As will be discussed in more detail in Chapter 4.2 ff. below, many evaluations were not well designed from a causal analysis point of view. None used an experimental design or a time series design in which data are collected at several points in time before and after an intervention. Control groups were used in only a few cases, and they were not randomly selected. In most of the cases, impact assessments were based on ex-post perceptions of changes in outcomes among persons interviewed. Such information is relevant and useful, but for obvious reasons it is not sufficient to establish the extent to which change has actually occurred.

Although few of the sample evaluations were able to provide detailed answers to the questions about impact and effectiveness, many explained quite well why they could not do so. The evaluation of a project in support of the Office of the Vietnam National Assembly (SE 02/12) is a case in point. One of the key questions in this evaluation concerned the impact of Swedish support to the strengthening of the National Assembly and the democratic process in Vietnam. In its description of recent changes in the Vietnamese political system and the increased transparency of the activities of the National Assembly, the report refrained from naively attributing those changes to Swedish support. Instead it argued more modestly that the Swedish intervention had stimulated the processes in question by providing know-how. It would not go any further than this. While trying to assess the likely effects of each one of the project components, it admitted frankly that, *"the extent of contribution of the project cannot be measured exactly"* (p. 38).

The following are variations on the same theme:

> "To analyse the results of this programme in terms of effectiveness and efficiency is not an easy task for a number of reasons: The programme's activities cover a whole range of aspects both related to human resource matters, training and general policing. Hence, there are difficulties related to the size and scope of the exercise especially on effect and impact level. The programme is but a minor contribution of other donor support as well as compared to the total cost. On a methodological level effects of police

activities on society is a very complex issue and the external factors that affect outcomes are many, for example unemployment, immigration, cultural values, etc. There are also complex issues related to statistics on crime where reporting methods, degree of reporting, etc. may vary over the years and between types of crime." (SE 05/14. Review of the Development Cooperation Programme between the South African Police Service and the Swedish National Police Board)

"…a word of caution is in order about causality. The objective of this evaluation is to assess the impact of specific Sida-funded activities. In all of the above positive tendencies, many different actors are involved and Sida/UNICEF plays just one part which, in many cases, cannot be distinguished from the rest… Documenting impact will often have to answer the question of attribution, i.e. to what extent a development intervention has contributed to attaining the goal and purpose. Impact is often assessed after the intervention has been completed. Nevertheless, it is the experience of many donors that impact studies must be planned before a given intervention is initiated." (SE 03/41. Sida funded Projects through UNICEF-Bolivia, 1989–2002)

"Impact is normally addressed through carefully designed field studies in the context of which the programme operates or has operated. Even under the best of conditions this is a difficult task, not the least because of difficulties of relating programme interventions to changes in the context in a manner of cause and effect. In this case where field investigations were ruled out (Somalia and Congo) it must be stressed that the impact assessment becomes very much a question of guesstimates, of informed speculations on the likely outcomes and lasting effects. Lacking both primary field data, focal studies and monitoring reports should not be construed or read as an impact assessment in any real sense of the word. Insofar as we have anything to say on the situation on the ground it is through hearsay and interviews with previous staff. Instead of recording footprints, what we can offer is a discussion of presumed footprints." (SE 04/36. Life and Peace Institute's Projects in Somalia and the Democratic Republic of Congo)

Statements like these in one report after another raise the suspicion that there is something wrong with Sida's evaluation system as a whole. How can there be such a mismatch between questions and answers? Why does Sida not get the requested information? This report does not attempt to provide a complete answer to this important question. As suggested in Chapter 6.2 below, however, we would not put all the blame on the evaluators. In our estimation the evaluators are often doing a reasonably good job, given the constraints

under which they work. Inadequate evaluation budgets could be part of the problem, and beyond this there is a variety of evaluability problems. As pointed out in the quotations above, technical problems sometimes stand in the way of a satisfactory assessment of effectiveness and impact.

3.4 Analysis of Implementation

Although this study was primarily intended to assess the quality of the results information contained in sample reports, we also looked at how the reports dealt with questions about implementation. To simplify the assessment we formulated six categories that cover the basic elements of any implementation process:

1. Leadership and governance
2. Planning
3. Financial management
4. Coordination
5. Networks and linkages
6. Organisational structures

A comprehensive analysis of implementation would normally contain views on all these issues – but it is quite possible that the TOR only focus on one or a few of them.

Table 7. Implementation analyses in the evaluation reports

	1	2	3	4	5	6	N/A	Total
Is there a plausible analysis of leadership and governance?	6	2	6	6	8	1	5	34
Is there a plausible analysis of planning?	1	3	2	15	5	1	7	34
Is there a plausible analysis of financial management?	10	4	2	4	6	1	7	34
Is there a plausible analysis of coordination?	5	5	2	10	8	1	3	34
Is there a plausible analysis of networks and linkages?	4	3	3	7	10	1	6	34
Is there a plausible analysis of organisational structures?	2	5	1	12	8	1	5	34

Key to ratings: 1 – very poor (or not done at all), 2 – significant problems, 3 – not quite adequate, 4 – minimally adequate, 5 – adequate, 6 – excellent, NA – not applicable, the question was irrelevant to that evaluation, or the issue could not be assessed because of a lack of information.

Source: The authors' assessment of 34 evaluation reports

In this section we look at the reliability of the analyses of various aspects of implementation. In our assessment we have used the word "reliable" rather than "accurate" to reflect our somewhat less rigorous way of conducting the analysis. It must also be noted that in a large number of reports it was not relevant for the teams to discuss leadership and governance (5), planning (7) and financial management (7) as the TOR did not include questions pertaining to such issues. It was also a problem that some of the terms, for example, governance, planning and network linkages are open to interpretation and are assessed differently by different authors.

Table 7 indicates that the analysis of development aid (financial) management is at times rather weak. There are several reports with significant problems and very few excellent examples. The analyses of organisational structures and network linkages obtain the best average ratings.

The background and expertise of the evaluators also appear to have influenced which aspects the evaluations focused on. A management consultant will be more interested in implementation processes and specific management issues than a technical expert who is likely to look more carefully at issues such as the design and results of the intervention. This makes it difficult for a small evaluation team to provide an equally solid achievement analysis for all the evaluation criteria and the various aspects of aid implementation discussed here.

3.4.1 Leadership and Governance

It is notable that leadership and governance issues – meaning in-depth analysis of the role(s) of the leader and top management in the preparation and implementation of an aid effort – are often left out of, or only marginally discussed in, the evaluations. An understanding of the importance of dynamics between individuals is most often missing, despite the emphasis in management and organisational research on the importance of individuals as champions and leaders of change processes.

The evaluation of the Institutional Development Programme at the Ministry of Education in Mozambique (03/29) is an example of a very good analysis of leadership and governance issues – it provides the reader with an increased understanding of complex processes combining individual and systemic factors. The main objective of the programme was to develop the capacity of the Ministry at all levels to manage the national education system in a way that supports the delivery of the Education Sector Strategic Plan and ensures an efficient and effective use of its resources. The evaluation report concludes that the programme failed to produce the planned results and the capacity development at the higher organisational and institutional levels. We find it interesting and of high quality for several reasons:

- A broad range of constraints is taken into consideration when answering a complex question. The complexity of the issue is accepted and properly addressed.
- Both internal and external constraints are identified within the Ministry of Education itself, in programme execution, but also in donor behaviour.
- There is a combination of systemic constraints (unclear roles, missing strategic framework, etc.) and a clear understanding of the role of leadership and personal commitment for effective implementation. It is also one of the few evaluations in which there is specific reference to a gender dimension.
- There is a separate and in-depth discussion of the role and effectiveness of the management adviser – illustrating some of the dilemmas and tensions in providing technical assistance.
- There is an understanding of the constraints represented by the organisational culture and structure of a ministry in a developing country like Mozambique:

  > "Public administration culture is heavily vertically hierarchical and authority is highly recognized by staff. This makes the staff strongly dependent on their bosses and closes the door to innovation….As to decision-making, this is heavily formally centralized and the delegation of competences is set by decree or dispatch, from one head of unit to the next, downwards on the hierarchy ladder." (p. 21)

Box 5. Example: Constraints in institutional development

"Weak and unclear role of the working group for institutional development

…According to the TOR the working groups are only consultative organs and have a technical nature. Their function is to give advice and prepare proposals for decisions to be taken…. This arrangement, to hold the working group responsible for planning, but not for implementation, has led to limited flow of information, little engagement between the plan and its implementation, the abandonment of activities and the low dissemination of the programme.

"New political leadership created new circumstances

Another important factor, which probably has affected the ownership and engagement in the programme, is that six months after the effective start of the programme a new Minister, Vice-Minister and a new Permanent Secretary were appointed….

"Lack of commitment and engagement in implementing the programme

The representatives of the working groups identified a lack of leadership and coordination of the programme as well as limited engagement from the Ministry in the programme. This lack of interest in taking forward initiatives has led to slow decision making processes and that many activities have not taken place… because of limited

delegation, limited internal communication, lack of incentives, limited capacity etc. These factors in turn are an effect of a hierarchical management tradition and organizational culture.

"Activities turned out less effective due to a missing strategic framework

Many activities turned out to be ineffective due to a lack of a strategic framework or objectives for the specific activities. Seminars and discussions were held without having defined how to use or take responsibility for actions… Members of staff have been trained in human resource development and in English, but there has been no follow up or analysis of whether staff have had the opportunity to practice this new knowledge.

"Reluctance towards recruitment of technical assistance

In the project document as well as in the annual plans technical assistance has been considered… but the national directors have been reluctant to recruit technical assistance within the programme. … The Team's conclusion is that the fact that W[orking]G[roup for]I[nstitutional]D[evelopment] presented TORs, not based on an articulated need in the organization, contributed to the reluctance.

"Gender awareness is lacking in the implementation of the programme

There is hardly any awareness regarding gender in the documentation of the programme and the team could not identify any activity aimed at strengthening gender awareness within the programme.

"Donor involvement in Working Group for Institutional Development

The representation of donor members in different working groups is an attempt to exert an increased influence, and to speed up the process of change. However, the impact of such donor involvement may be counter-productive in taking both ownership and responsibility away from the Ministry. ….

"Lack of systematic monitoring

A systematic follow-up and monitoring was never implemented which has also contributed to limited impact of the programme and its lack of cost-effectiveness. The annual work plans were of very poor quality and very limited analysis regarding the failure to implement strategic activities was made. If a proper monitoring of results had been made, measures could have been taken to adjust the programme or to stop disbursement unless strategic institutional development initiatives were taken."

Source: SE 03/29. Institutional Development Programme (RCI) at the Ministry of Education in Mozambique p. 24 f.

3.4.2 Planning

It is not uncommon for evaluations to point to shortcomings in planning, for example, that the plans were not flexible, that there was no room for contingencies or that they were deficient in some other way. The problem is that the evaluations do not show exactly what led to the lack of flexibility. They lack the kind of concrete discussion seen in the example above, which is necessary for understanding what really went wrong.

The evaluation of the State Financial Management Project in Mozambique (04/29), where Sida had supported the Ministry of Planning and Finance for

15 years, aims to draw conclusions on the approach of the project and its impact. The report offers an interesting analysis of the project planning process, assessing both its substance and terminology. The evaluation team stated that *"the plan of operation, particularly for the early phases, is less than effective as a plan"* (p. 24).

> "A process approach does not mean that no or limited design takes place. It is used more because of uncertainty about interventions needed to achieve what is proposed. In any event, the first stage of any project (including process approach projects) should be planned in detail.
>
> "The lack of clarity in use of terms, the repeat of outputs and activities in years 1 and 2 and the one to one relationship between output and activity, make for less than rigorous planning and impaired monitoring and evaluation." (p. 25)

The report concludes wisely with a plea for a robust project plan, but not necessarily a specific planning model:

> "The later plans of operation are progressively more precisely defined, in terms of time and results. However, although the logical framework approach was introduced to the Sida project planning arrangements from 1995, there is no evidence of the log frame being used as the principal planning and monitoring tool. There are those that argue that such a planning tool should not be imposed on recipients, as that would be non-participatory and that using it requires special skills and higher level linguistic ability. Nevertheless, whatever tool is used, a robust project plan is crucial." (p. 25)

3.4.3 Financial management

An analysis of financial management is in many cases an essential part of an evaluation of the implementation process. The evaluation of the Mozambique State Financial Management Project (04/29) provides a reliable analysis of financial management, which is in fact the main purpose of the evaluation. The Management Audit of the Swedish Red Cross (04/04) includes auditors in the team of evaluators. The evaluation of the Institutional Development Programme at the Ministry of Education in Mozambique (03/29) analyses not only the public financial management system, but also the wider role and functions of such a system in the government. The problem is not only the system itself, but also how it functions in the government and with external donors. The dynamics of a public financial management system are well analysed and explained. But there are also a few other cases with a satisfactory analysis of financial management, for example:

Box 6. Example: Analysing state financial management

"The development of a public financial management system has created tensions within both the Government and the development partner community. The present model of public accounting dates back to 1881 and is essentially a cash based system in which budget releases are provided on a rolling interest basis… This is incompatible with a modern budget drive system in which funds are drawn down on the basis of activity based plans translated into cash flow terms... The other major problem with the existing system is that its coverage is only partial… a substantial amount of funds are essentially off-account. This in turn created major issues concerning misappropriation and leakages of funds, which are of particular concern to the development partner community and the IMF in particular… The extremely limited capacity of public financial management expertise in Mozambique has created problems about how to resolve this problem. The MPF has favoured a top down 'single size' fits all solution. However …"

Source: SE 04/29. Mozambique State Financial Management Project

When the evaluation of financial management is perceived as only a by-product of the evaluation, it is often much weaker and superficial, covering selected financial issues but not necessarily financial management as such. The evaluation of Sida Support to the University of Eduardo Mondlane in Mozambique (03/35) considers financial management in a separate section, discussing delays in disbursements, underspending of resources, etc., but contributes less to an understanding of the financial system and how it can be improved. The evaluation of the Zimbabwe National Network of People Living with HIV/AIDS (03/05) is an organisational assessment but devotes only two pages to a description of certain technical aspects of financial management. A comprehensive analysis of financial management requires special skills – technical knowledge of the system itself combined with an understanding of its interplay with the institutional context.

3.4.4 Coordination and networks

Coordination and networking are recurrent issues in implementation analysis. The evaluation of Integrating Natural Resource Management Capacity in South East Asia (05/13) covers five countries and is a regional network initiative to promote new agro-forestry policies and practice; as such it is an evaluation of a network. The evaluation of the Zimbabwe AIDS Network (04/10) is also an evaluation of a network, but it does not discuss the characteristics of a network organisation.

The evaluation of Research Cooperation between Vietnam and Sweden (02/06) presents an analytical framework for institutional capacity development, bringing in a network perspective. The framework distinguishes between:

> *"Human resource development* – which is concerned with how people are educated and trained, how knowledge and skills are transferred to individuals, competence built up and people prepared for their current and future careers…
>
> *Organizational development* – which seeks to change and strengthen management systems in specific organizations in order to improve performance…
>
> *Systems development* which is a broader concept – including the linkages between the organizations, and the context and environment within which organizations operate and interact. In respect of the Swedish-Vietnamese research cooperation, it is particularly important to distinguish between network and linkages among organizations, which include the network and contact between organizations that facilitate or constrain the achievement of particular tasks." (pp. 23–26)

The strength of such a framework is that it opens up to and supports the analysis of interactions between various levels. It also explains the role and significance of external networks and linkages between micro and macro processes.

One of the most common comments on management issues is that coordination has been weak. Almost all evaluations have some conclusion to that effect. However, they seldom specify what was wrong, and whether it was the end product that was poorly coordinated or whether it was the process that was weakly designed. We did not see any analysis of which means of coordination had been used, how expensive coordination was, or whether more cost-efficient approaches to coordination could be conceived.

3.4.5 Organisational structures

In the main, the evaluations do not look at organisational structures, though this could be a bias in our sample. Few evaluations study organisations as such. The object of most evaluations is a project or programme, and it may not have been obvious how organisational structures should be assessed. None of the evaluations contains an analysis of whether the structures per se could be improved, where such aspects of organisational structure as span of control, division of labour and levels of decentralisation could be improved or the merits of a chosen design.

The Management Audit of the Swedish Red Cross (04/04) is an assessment of the Red Cross structure in Sweden and internationally. It provides a systematic and reliable, albeit limited, analysis of organisational structures and financial management. The main focus is on internal systems and procedures for planning and implementing projects. It covers the Red Cross's international network, but does not explain to what extent Red Cross perform-

ance is enhanced and/or constrained through interaction with external international networks.

The evaluation of the Institutional Development Programme at the Ministry of Education in Mozambique, on the other hand, is an example of a comprehensive, in-depth analysis of organisational capacity for implementing the programme. Points of particular significance:

- The report looks at organisational capacity at three levels (national, provincial and district).
- The broad concept of capacity is broken down into relevant components that are analysed separately.
- There is a dynamic process perspective in the analysis explaining who the actors are, what the important processes are, the constraints, etc.
- Internal and external aspects of organisational capacity are discussed – human resource management within the Ministry, as well as external public sector reform constraints, which are beyond the Ministry's control.

Box 7. Best practice: Assessing organisational capacity

Organizational Capacity at National, Provincial and District Level

Human resource management

… The human resource management is done at central level, what makes it extremely difficult to avoid delays when contracting teachers… no institutional development reform can be applied without taking this aspect into consideration, since it is one of the serious bottlenecks of the sector.

Centralization and strategic management

The formal centralization of decision-making calls the planning procedures to the central level, making it rather difficult for the lower levels… to have access to the planning know-how… The competence for a "sector-wide approach", which exists to a certain extent at central level, does not consistently replicate itself at the successive levels down to the school.

Reform constraints

A number of constraints can be identified which affect the degree to which reforms of institutional development character can be applied. First, there is the legal country framework… Low salaries are also a well-known constraint… Public administration structures are heavily vertically oriented…

Professional qualification

In spite of the generally good academic qualifications of most of the MINED staff, they often do not possess professional qualifications that give them the expertise for technical work, since most of them have been teachers.

Source: SE 03/29. Institutional Development Programme (RCI) at the Ministry of Education in Mozambique

4 Methods and Evidence

4.1 Where is the Evidence and How is it Used?

There is no doubt that most evaluators generate a lot of evidence through interviews, observation and other methods of data collection. We assessed the extent to which the reports presented empirical material, whether the analysis was exhaustive and if the findings and conclusions were supported by the empirical data. Generally speaking, the reports did not provide sufficient empirical evidence. The ratings are shown in Table 8.

Table 8. The empirical basis for analysis

	1	2	3	4	5	6	N/A	Total
Does the evaluation present empirical material in the report?	0	3	3	17	8	3	0	34
Is the analysis relating to the evaluation questions exhaustive?	1	3	6	12	10	2	0	34
Are findings and conclusions supported by the data?	1	3	5	11	11	2	1	34

Key to ratings: 1 – very poor (or not done at all), 2 – significant problems, 3 – not quite adequate, 4 – minimally adequate, 5 – adequate, 6 – excellent, NA – not applicable, the question was irrelevant to that evaluation, or the issue could not be assessed because of a lack of information.

Source: The authors' assessment of 34 evaluation reports

On the first question, "Does the evaluation present empirical material in the report?", more than half of the reports are rated as not quite adequate or minimally adequate – there are a few with very weak empirical evidence, as well as a few very good cases. There is clearly a need for more and better empirical evidence and systematic use of such information in a majority of the reports. The same is true for the second question, analysis of evidence in relation to the evaluation questions: the analysis is not sufficiently exhaustive in most of the reports.

When it comes to supporting findings and conclusions with data, 24 out of the 34 reports were rated as minimally adequate. Under "minimally adequate", however, we included reports in which most of the evidence supported the conclusions, even though other evidence might point towards a

different interpretation. We deemed an evaluation to be "minimally adequate" when the major conclusions seemed to be supported by data, although they could still be questioned. If we gave a 5 or 6 the conclusions were less open to questioning. The ratings suggest that the empirical data provided strong support for the conclusions in only 13 of the 34 reports.

Evaluations are supposed to use scientific research methods but will often have to compromise on the application of such methods because of limitations in time and resources. It is important to acknowledge the difference between evaluations of the kind discussed in this report and academic research, but evaluations still need to satisfy two important requirements: data and information should be collected systematically, and conclusions should be based on solid evidence, otherwise there is a risk that evaluations are reduced to, and perceived as, only subjective opinions.

Hence, it is a weakness that most evaluations tend to use a narrative and descriptive form in the analysis without drawing upon empirical evidence. Evidence is not systematically presented, utilised and integrated into the analysis. Findings and conclusions are thus characterised by broad sweeping statements – based on impressions gained through the evaluation process. A typical example of this is:

> "The overall impression gained from the 43 visits to local civil society organizations is that the work the Swedish and local civil society organizations are doing is important and effective in that it produces results in line with Sida's overriding development goals." (05/16 p. 40)

There are a few examples of reports with almost no empirical data. They appear as subjective testimonies by a team "looking at" an activity and expressing their own opinions, without giving the reader the chance to assess the reliability of the findings. The evaluation of Sida's Work Related to Sexual and Reproductive Health and Rights (04/14), for example, included visits to several countries and international organisations, but from reading the report it is difficult to get an understanding of what the evaluators observed and learned from those visits. There is a general and narrative text analysing broad trends and policies, but the text appears weakly anchored in empirical material, i.e. in what the evaluators saw and heard during the visits. The country findings are not clearly reflected and presented in the report.

In contrast, there are good reports with a lot of empirical material and exhaustive analysis. The evaluation of Sida's Health Support to Angola (03/19), for example, provides a relatively brief but concise description of the programme context and the various components of the programme. For each component, major achievements and constraints are discussed using a combination of statistical data, observations from site visits and information from interviews.

Box 8. A good example: Presentations of facts and findings

The evaluation of Sida's health support to Angola assesses all components of the programme in the same systematic manner in the Findings chapter, for example:

3.1. Maternal health
3.1.1. Objectives, purposes and results – planned and achieved
3.1.1.1. A review of plans and reports and indicators
3.1.1.2. Field evidence
3.1.2. Changes during the period
3.1.2.1. Suggestions from the 1999 evaluation
3.1.2.2. Major achievements and major constraints

At the end of each sub-chapter there is a summary, for example:

> "The most impressive breakthrough in relation to maternal health care in Angola is the decen¬trali¬zation of institutional maternal care. There are now 33 antenatal clinics in Luanda and the number of peripheral delivery wards has increased to 15. In 1999, the reported number of institutional deliveries at peripheral clinics was 55.992 compared to 82.250 in 2002. The program has contributed by making significant investments, not only in building and equipment, but also in staff training.
>
> The corresponding information from the referral hospitals are presented in the table below… The MMR continue to be high and at the same high level registered 1989 when the programme was initiated… The peripheral maternity units are under-utilized… the capacity is more than double… The informal fee system may be a reason for the low utilization of the maternal health care."

This is a simple and straightforward presentation, which consists of important basic elements:

– Statistical data and an independent assessment of the validity and reliability of the data
– Utilisation of evidence collected during visits to clinics and interviews
– Efforts to explain findings based on field experience
– Assessment of plans and intentions – comparing targets with results

A similar structured approach is followed in the assessment of all of the programme components.

Source: SE 03/19. Sida's Health Support to Angola 2000–2002 p. 10 ff.

4.2 The Design of Evaluations

In principle, an evaluator can choose between various designs to study impact and effectiveness. Six alternatives were included in our scheme of analysis: randomised control groups pre- and post-test design (classic experiment), non-randomised groups pre- and post-test design (quasi-experimental design), one-group pre- and post-test design, one-group time series design, judgemental sample and case study design, and narrative analysis. The majority (67%) of the reports used only narrative analysis, based on a review of available documents and information from interviews. Some combined nar-

rative analysis with a case study approach. Although most of the evaluations could be called case studies, as they describe and assess one (and not two or more) development project or programme, we decided to use a stricter definition of what constitutes a case study design: a systematic and analytical approach.

Table 9. Designs chosen in the evaluations

Design alternative	Number	Percentage
Randomised control group pre-test – post-test design	0	0
Non-randomised groups pre-test – post-test design	0	0
One-group pre-test – post-test design	0	0
One-group time series design	0	0
Judgmental sample, case study design	12	33
Narrative analysis	22	67

Source: The authors' assessment of 34 evaluation reports

It is striking that none of the more "complex" designs were used in the evaluations in our sample – not even the time series or quasi-experimental designs. In some academic circles the randomised trial design is considered the only "scientific" approach and in some countries an experimental approach with a control group is required in all evaluations (e.g. the USA, the UK for the education sector). There is a strong movement towards experimental design with randomised control groups in the so-called evidence-based approach to evaluation. There is thus likely to be a requirement for more variety and more attention to design in the future. Box 9 contains a list of different designs, with some explanations.

Box 9. Examples: Evaluation designs

Experiments with Randomized Controls: Outcome measures among targets to whom an intervention is given in a provisional tryout before the permanent intervention are compared to outcome measures among an equivalent group, created through randomization—randomized controls—from which the intervention is withheld or which has been exposed to other intervention(s) (classic experiments).

Experiments with Matched Controls: Outcome measures among targets to whom a provisional tryout is given or who has been exposed to the permanent intervention are compared to outcome measures among a theoretically equivalent group, created non randomly through matching—matched controls—from which the intervention is withheld or which has been exposed to other intervention(s) (quasi–experiments).

Generic Controls: Outcomes of the permanent intervention among targets are compared with actual out¬comes or estab¬lished norms about typical outcomes occurring in the equivalent larger population not covered by the intervention.

Reflexive Controls: Data on outcome dimensions among targets who receive or have received the permanent intervention are compared to data on the same outcome dimensions among the same targets, as measured before the intervention.

Statistical Controls: Outcome changes among participant and non participant targets of the permanent intervention are compared, statistically holding constant differences between participants and non partici¬pants. Statistical controls are also applicable to full coverage interventions.

Shadow Controls: Outcomes among targets who receive or have received the permanent intervention are com¬pared to the judgements of experts, program managers, staff, or participants on what outcomes they believe would have happened without the intervention.

Case Study (Process Tracing, Process Evaluation): To find out the extent to which the intervention has influenced outcomes, the intervention formation, the intervention imple¬mentation, the addressee response, the organization of the control function, the actions of the principals after the adoption of the intervention, and the intervention context is studied as a rich case in its natural surroundings in order to discover and establish explanatory factors besides the intervention.

Source: Vedung (1997)

There could be several reasons for the choice of design. Many evaluators may not be conversant with the panoply of possible designs and might not have considered the alternatives. The more advanced designs may have been assessed but judged as inappropriate given the terms of the assignment. For instance, it is difficult to envisage an experimental design for the evaluation of Sida's support to the Vietnam National Assembly. There are of course also limitations on how Sida evaluations are carried out in terms of time and available resources.

An evaluation usually includes some time for preparatory work at home for the consultant, one or two weeks of visits to partner countries and some time for report writing. More time and several consultants are involved in large evaluations, but the approach is more or less the same. Time series designs require similar data to be collected for at least two different points in time – often requiring more than one visit to a country with significant intervals. An experimental design requires the evaluation team to find at least two comparable geographical areas and to be allowed to increase the evaluation budget in order to collect and compare data from various sites, which is not viable in many cases.

If Sida wants more and better information about impact, such information will either have to come from improved impact monitoring systems or through more long-term evaluations that are designed specifically to collect data and information about change over time. Improving the design of evaluations will require more resources and better planning however. An evaluation based on experimental or quasi-experimental methods can usually not

be initiated after the project has been implemented, but has to be included in the project design and begin at the same time as the project itself. The challenge for Sida is how to support the utilisation of more demanding designs (in terms of increased time and costs) in future evaluations.

4.3 Data Collection

Looking at the choice of methods for data collection, the picture is slightly more varied. The dominant pattern, however, is clear: the two most common data collection methods are document analysis and, to a lesser extent, open-ended interviews including ad-hoc observation (i.e. visits to one or more project sites). Only two evaluations used standardised interviews and structured observation. Nine used focus group interviews and five of the reports used surveys. This means that questionnaires and interview guidelines are not attached to many of the reports, for the simple reason that they do not exist. Most of the interviews were open-ended or semi-structured.

Some reports emphasise the virtues of methodological triangulation – using several methods to answer and shed light on various aspects of the same question. The figures in Table 10 seem to indicate a low level of triangulation: most of the evaluation processes were organised in a similar pattern – document analysis followed by open-ended and semi-structured interviews and, if relevant, visits to the respective project site. The evaluation of the Development Cooperation Programme between the South African Police Service and the Swedish National Police Board (05/14) is a typical example of triangulation.

Box 10. Example: Methodological triangulation

"The review has been carried out through a study of project documents, plans, reports, agreements and financial data, as well as general documents related to the Structural Adjustment Programs in South Africa. In addition a number of interviews have been made with the Structural Adjustment Program and project managers in both countries. The interviews were combined with site visits in Gauteng, Northern Cape, Free State and KwaZulu/Natal.

"In the evaluation of a UNICEF project in Bolivia (03/41) a multi-dimensional understanding of poverty was developed comprising of (a) basic needs, (b) livelihood, (c) resources and vulnerability, (d) social and political deprivation, and (e) psychological deprivation. The problems of measuring impact directly using quantitative data was recognized, but several data collections were used: (a) Community visits beginning with the construction of a timeline, (b) focus group interviews in each community using an Impact Assessment Matrix, (c) case study interviews, (d) and lastly relevant national statistics."

Source: SE 05/14 The Cooperation between the South African Police Service and the Swedish National Police Board

Most of the evaluated initiatives are relatively large – covering a sizeable geographical area and/or target group. It is therefore not surprising that only one evaluation gathered data from the whole population. It is more striking that a majority of the reports (85%) gathered data from only a purposive or purely ad hoc sample. In other words, the typical sample of informants was selected based on the evaluators' (and to some extent Sida's and the recipients') own decisions. During the course of an evaluation, additional informants were interviewed ad hoc. The same sampling pattern seemed to apply to the document reviews.

Given the limited time, the evaluators collected as many reports as possible from Sida and the project/organisation at the beginning of the process. In a few cases the TOR actually asked for a more systematic literature search; otherwise the evaluators tended to rely on the literature and documents most easily available. In the evaluation of the Water Education in African Cities (04/21) the first part of the assignment was devoted to a comprehensive review of written documents, but this is a rare example.

Few evaluations made an effort to collect data systematically from a random sample. This might be explained by the absence of experimental designs mentioned earlier, or it could be related to the time constraints involved. Whether it is for a questionnaire, focus group, observation or document analysis, the evaluators have to select a sample (unless of course they choose to address everyone). The choice of sample is very important, but only two evaluation reports contain any discussion of the principles they applied and how the selection affected the findings.

Table 10. Data collection methods in the evaluations

Data collection alternative	Number	Percentage
Surveys	5	15
Focus group interviews	9	26
Individual standardised interviews	2	6
Individual structured interviews	7	21
Individual open interviews	32	94
Structured observation	2	6
Ad hoc observation	12	35
Document analysis	31	94

Source: The authors' assessment of 34 evaluation reports

Once the data collection methods have been chosen, the work of constructing the instrument for data collection can start. There are many ways to compile a questionnaire, for instance. Should the questions be open or closed? What should the mix of questions be? How many questions should there be? If you ask for opinions or values, what type of scale should you use? There are also many questions that can (and should) be posed regarding in-

terviewing: how is a focus group organised? Where, and with what agenda? All these issues must be assessed.

For an evaluation report to appear reliable we need information on the instruments and procedures for data collection. We refer not only to questionnaire formats, interview guidelines and the like, but also to indicators, rating scales and benchmarks. We did not find sufficient information in the reports on the way indicators had been defined. There were a few evaluations, however, in which indicators had been defined at the outset of the evaluation and then used deliberately and systematically to collect data about performance. Rating scales and benchmarking were used only in a small number of reports.

A good example of the systematic use of indicators is found in the evaluation of Contract-Financed Technical Cooperation and Local Ownership (03/09:1). The report describes how the evaluation process moved through a number of steps:

- "First, a number of characteristics and dimensions with which to characterize contract-financed technical cooperation and local ownership were identified.
- Then each characteristic and dimension was given an operational dimension which in turn allowed the definition of indicators and scales. These were then used in each project to characterize, both the application of contract-financed technical cooperation and local ownership." (pp. 6–15)

The main indicators for the various characteristics and types of ownership in each of the different projects were presented and analysed at the end of the report. This made it possible to present a considerable volume of information in a concise and systematic form and to facilitate comparison between projects as well as countries.

Table 11. Deployment of instruments for data analysis

Instrument	Number	Percentage
Qualitative indicators	18	55
Quantitative indicators	11	33
Rating scales	8	24
Benchmarks	6	18

Source: The authors' assessment of 34 evaluation reports

4.4. Assessing Design and Methodological Choice

Almost all the reports contained a chapter on methodology that described the adopted methods and how the evaluation teams had applied them. These accounts were of variable quality. The majority of the reports described the methodology briefly but did not point with care to its limitations. In ten of the reports there was no discussion of threats to reliability and validity, or other comments on the quality of the findings and conclusions. It is surprising that the evaluators did not care to make the reader aware of potential weaknesses and limitations, as this could protect them against unfair criticism. On the whole, much remains to be done with regard to methodological transparency.

It was difficult to assess the quality of the data collection instruments as, in some cases, they were not even described in the report. The lists of questions for open-ended and structured interviews and survey instruments are only attached to a few reports. Where the reports do present their data collection instruments, benchmarks, rating scales, etc., they are mostly well constructed, appropriate and relevant. The evaluation of a regional training programme in Sri Lanka (04/18), for example, carried out a simple e-mail survey, which appears to provide useful feedback on course content and implementation.

We found that almost all the evaluations followed the same design and used similar methods to collect information and answer questions – a combination of document review and interviews. There is very little methodological variety in our sample of evaluations. From one point of view this can be regarded as a weakness. Nevertheless, we rated the choice of design to be appropriate in the majority of reports (26 reports rated in categories 4 to 6), though we were slightly less convinced about the appropriateness of the data collection methods (28 reports in categories 3 to 5).

Table 12. Assessment of methodological choices

	1	2	3	4	5	6	N/A	Total
Description of Methodological Choices								
Is there a section that describes the methodological choices fully? [7]	2	7	6	8	8	3	0	34
Is there a discussion of threats to reliability and validity?	19	1	2	10	1	1	0	34

[7] This question can be said to consist of two or maybe three sub-questions: (a) is there a separate section describing methodological choices, (b) is that section reasonably exhaustive, and (c) were the described choices well argued? This may be true also for other questions and the original battery of questions will have to be revised for later use.

	1	2	3	4	5	6	N/A	Total
Description of Methodological Choices								
Can the reader make an independent assessment of the methodology?	4	1	8	9	10	2	0	34
Is there a clear statement of limitations to the evaluation?	11	4	5	7	4	3	0	34
Designs for Causal Analysis								
Is the design of the evaluation appropriate, given constraints on budget, timing, and preparatory work?	0	2	5	9	15	2	1	34
Data Collection Methods								
Are the data collection methods chosen appropriate to answer the evaluation questions?	2	3	5	12	11	1	0	34
Is there a relevant and adequate selection of sources of data?	0	4	6	11	12	1	0	34
Does the choice of methods suggest that the evaluation will obtain reliable and valid data?	2	3	8	11	8	1	1	34
Instruments for Data Collection and Analysis								
Are the instruments for data collection well constructed?	10	1	2	0	5	0	16	34
Are indicators appropriate?	3	4	1	5	8	0	13	34
Are benchmarks fair and relevant?	12	0	2	2	2	0	16	34
Are rating scales well designed?	10	1	2	0	4	0	17	34

Key to ratings: 1 – very poor (or not done at all), 2 – significant problems, 3 – not quite adequate, 4 – minimally adequate, 5 – adequate, 6 – excellent, NA – not applicable, the question was irrelevant to that evaluation, or the issue could not be assessed because of a lack of information.

Source: The authors' assessment of 34 evaluation reports

Note however, that "appropriate" here means "realistic", given the available time and resources, but not necessarily ideal from a research perspective. There is a drift towards certain choices, leaving out more complex and demanding designs and methods for very practical and pragmatic reasons. The designs and methods are not necessarily the most desirable, but they are manageable and realistic and in that sense appropriate. They represent a

pragmatic compromise with the ideal requirements, which could unfortunately undermine the quality of the evaluation. It should also be noted that the terms of reference often prescribe the choice of data collection methods. This, in combination with limited time and resources, often ruled out more complex designs and methods. Thus, in some cases Sida rather than the consultant was responsible for the methodological shortcomings.

We have to conclude that the selection of methods and data collection is not adequate – Table 12 shows that most ratings fall around the minimally adequate line. Only one evaluation report could be placed in category 6 and two were placed in category 1, with the majority (19) in the middle categories (3 and 4).

5 Conclusions and Making Recommendations

5.1 Introduction

Sida's evaluation manual states that evaluation in development cooperation serves two general purposes: accountability and learning (Sida 2007, p. 12 ff.). Accountability is achieved when the evaluators report back on implementation and results of a Sida-funded activity, where responsibility has been delegated to an implementing counterpart. To fulfil the purpose of learning, an evaluation "is expected to produce substantive ideas on how to improve the reviewed activity or similar activities" (Sida 2007, p. 17). Evaluations therefore need to be transparent, consistent and reliable, as discussed in previous chapters, and to formulate clear recommendations and lessons to be learned. The OECD/DAC Working Group on Evaluation has formulated evaluation standards, two of which relate to conclusions, recommendations and lessons learned:

> *"9.1 Formulation of evaluation findings.* The evaluation findings are relevant to the object being evaluated and the purpose of the evaluation. The results should follow clearly from the evaluation questions and analysis of data, showing a clear line of evidence to support the conclusions. Any discrepancies between the planned and actual implementation of the object being evaluated are explained..."
>
> *"9.3 Recommendations and lessons learned.* Recommendations and lessons learned are relevant, targeted to the intended users and actionable within the responsibilities of the users. Recommendations are actionable proposals and lessons learned are generalizations of conclusions applicable for wider use." (OECD/DAC 2007, p. 9)

In our assessment, we formulated eight questions based on these two standards. The results of the assessment are summarised in Table 13. We found that most evaluations in our sample respond to their TOR (see section 3.2.) and that their conclusions are clear and consistent. Most of them also provide recommendations that are anchored in the analysis and conclusions, although there is often a lack of empirical evidence, as we have shown above. These are very important quality criteria and they say much about the overall usefulness of the reports. However, when probing whether the evaluations had formulated useful lessons learned our findings were disappointing. We will return to the matter of lessons learned later in this chapter.

Table 13. Overall assessment of conclusions, recommendations and lessons learned

	1	2	3	4	5	6	N/A	Total
Are the conclusions in the evaluation clear and consistent?	0	4	2	12	13	3	0	34
Do the recommendations follow from the analysis and conclusions?	1	4	4	6	14	4	1	34
Are the recommendations practical; can they be translated into decisions?	1	2	5	10	10	5	1	34
Are there recommendations for clearly specified groups of actors?	5	1	5	9	9	4	1	34
Are there relevant, and for an informed audience, interesting lessons learned in a specific section?	17	2	5	5	4	1	0	34
Can an informed reader identify and make sense of lessons learned through the intervention?	1	4	11	8	8	2	0	34
Has the evaluation added to a general understanding of development cooperation?	1	8	10	8	7	0	0	34

Key to ratings: 1 – very poor (or not done at all), 2 – significant problems, 3 – not quite adequate, 4 – minimally adequate, 5 – adequate, 6 – excellent, NA – not applicable, the question was irrelevant to that evaluation, or the issue could not be assessed because of a lack of information.

Source: The authors' assessment of 34 evaluation reports

Even if useful recommendations and lessons are provided, however, for learning to occur as a result of evaluations, the organisations needs to receive these in a system that facilitates or enables learning and management responses to evaluations.

5.2 How Evaluation Reports Conclude

The overall rating of the evaluations in the sample is positive. The majority (28/34) of the reports were found to draw satisfactory conclusions. In 13 cases the conclusions were rated as quite good and 3 were considered excellent. Nonetheless, 6 evaluations did not pass the test with regard to the conclusions they offered.

The Joint Committee's Program Evaluation Standards include a "justified conclusions" standard: *"The conclusions reached in an evaluation should be explicitly justified, so that the stakeholders can assess them"* (Joint Committee on Standards 1994, A10). This means that the conclusions should be based on all the information collected and the evaluators should indicate what can be derived from the data, both in support of, and possibly against, the main conclusions.

Patton (1997, p. 307): suggests a framework for conclusions

1. *Description and analysis:* Describing and analyzing findings involve organizing raw data into a form that reveals basic patterns.
2. *Interpretation:* What do the results mean? What's the significance of the findings? Why did the findings turn out this way? What are possible explanations of the results? Interpretations go beyond data to add context, determine meaning, and tease out substantive significance based on deduction or inference.
3. *Judgement:* Values are added to analysis and interpretations. Determining merit or worth means resolving to what extent and what ways the results are positive or negative. What is good or bad, desirable or undesirable, in the outcomes? Have standards of desirability been met?

This framework shows how the process of arriving at conclusions draws on the empirical data and the analysis, as well as on interpretation and judgement. The conclusions themselves are the synthesis of this process – the end result of the presentation and the discussion, the consolidated statement of what the evaluation team has found. Factual conclusions should explicitly build on these steps and be clearly derived from them.

What does it mean by conclusions should be clear and consistent? It may seem difficult to pinpoint exactly what is meant by clarity and consistency. Yet conclusions are clear if they contain no obvious ambiguities and vagueness and are easy to understand, and they are consistent if they contain no contradictions. Box 11 presents an example of what we consider to be clear and consistent conclusions from the sample of reports.

Box 11. Example: Good practice – clear and consistent conclusions

6.1 Overall project relevance

To determine the project relevance in terms of the overall objective, the outcome rely on whether there is any democratic process to support, in the meaning of a development towards a more democratic society.

We have noted in chapter 4 that there are a number of ways in which the present system of government in Vietnam falls short of what, in Sweden and broadly by the international community, is seen to be fundamental to a functioning democracy as no real alternatives to the ruling party can be presented to the electorate. The voters are not free to elect their representatives of their own choice and no real alternatives to the ruling party policies can be presented to the electorate. There are also restrictions to the freedoms of opinion and expression and independent media do not exist. There is furthermore, as mentioned, no expressed intention of the leadership to change that situation. Therefore there is reason to question whether there is any "democratic process" in Vietnam to enhance.

In spite of these shortcomings we find there are reasons to conclude that there are possibilities to promote a democratic development in Vietnam. Even if there is no declared intention to systemic change, changes occur through the many reform processes currently taking place in Vietnam. And although relatively on a small scale, more transparency and publicity, a more open public debate, more focus on parliamentary supervision of the executive create a dynamism that obviously is seen as necessary and welcomed by many, if not necessarily by all. The liberalisation of the economy is an important driving factor.

These visible changes and ongoing reforms can be interpreted as a transition process, in the sense that we can observe changes also in the political procedures that determine the distribution of powers. The parliament has gained more formal powers through the Constitutional changes, and also through the new organisational structure that will increase the number of full time parliamentarians. This will certainly create a more efficient parliament with improved capacity in key areas such as law making and supervision. A coming new law on supervision would add to this development situation. There is no doubt the parliament plays a very active and important role in a democratisation process. Indeed, although division of power as conceptualised by Montesquieu is not part of the official ideology of Vietnam, the Parliament is increasingly assuming the role of a "peoples tribune" in the government structure, and a platform for political debate …

6.4 Mode of co-operation

The Mission has the impression that the project parties may have had different expectations on the role each party was supposed to play in the co-operation. From the side of Sida the project was perceived as an institutional co-operation of sister institutions where the two co-operating institutions would gradually develop the project content and deepen the co-operation from the level of exchange of experience to joint problem solving. The Riksdag Administration would develop a consulting role.

From the side of the Riksdag Administration, the role of facilitator of exchange between parliaments is familiar, whereas the role of acting as a consultant in joint

problem solving is unfamiliar and even questionable. Also from the side of ONA it seems as the expectations have been that the project should provide knowledge inputs to facilitate conceptualisation of development options in the agreed subject areas. A deeper involvement in long-term development activities by the Riksdag Administration does not seem to have been expected during this phase.

However, in a project of this nature the Mission would have expected the parties to review the project document thoroughly after the first year to analyse the achievements in relation to the objectives and to review the coherence and realism of planned outputs and activities. This has not been done as far as the Mission has observed. The absence of this kind of follow up indicates that the parties have regarded the project document as unchangeable. In a development project of this kind the experiences gained and the new circumstances that arise should be reflected in critical review of objectives outputs and activities and affect the plans made. The absence of this type of critical review may explain why certain outputs have remained whereas activities to achieve the output have been cancelled.

Another explanation why more long term development oriented activities – such as envisaged studies – not have been implemented may be found in a lack of readiness on the part of both parties to involve in such activities or that the more hands on activities, seminars and study tours, have fully absorbed the capacity of ONA and Riksdag Administration. If the latter explanation were valid it would indicate that there is a limit for the involvement of ONA staff and expertise from the Riksdag Administration in activities requiring active participation in several consecutive activities over a longer period of time.

Source: SE 02/12. Strengthening the Capacity of the Office of the Vietnam National Assembly, p. 38 ff.

We rated this report as being of high quality for a number of reasons. The report clearly distinguishes between Overall assessment (Chapter 6), Recommendations (Chapter 7), and Lessons learned (Chapter 8). The conclusions are clear – there is no doubt as to what the authors think – and they are also analytical, presenting arguments for and against. The writing is frank and addresses difficult issues head on.

Evaluations that rate lower in our assessment tend to present conclusions that lack sufficient data support. Consistency is also a problem in many cases. The object of inquiry, whether it is a project, programme or policy, is usually complex, and there are cases when the evidence points in different directions. Nonetheless, it is the task of an evaluation to make sense of the mess, while not oversimplifying it.

If the presentation of the data and analysis is transparent and comprehensive, it is easier to create clarity and consistency. The reader can then follow and assess the argument more easily and see how evaluative statements are grounded in comparisons between facts and value criteria. The more comprehensive the explanation of the analysis, the more likely it is that conclusions will emerge as clear and convincing. In some cases the evaluators pro-

ceed directly to summative results statements without describing how they arrived at them. Anyone who has interviewed large numbers of people knows that there are differences of opinion. Conclusions are not credible if this is not reflected in the analysis.

The reports that were given a high ranking on clear and consistent conclusions were all rather long, at 60 to 80 pages. This could be regarded as a problem, as many readers do not have much time to spend on reading reports. If, however, the report is well structured, has a clear executive summary, and otherwise helps the reader along, this may compensate for its length. Another possibility is to put some of the data in annexes.

5.3 Recommendations for Action

As stated in Table 14, we have used three main criteria to assess the quality of recommendations:

1. Do the recommendations follow from analysis and conclusions?
2. Are the recommendations practical: can they be translated into decisions?
3. Are the recommendations for clearly specified groups of actors?
4. Are the value judgements from which the recommendations follow clearly stated?

The majority of the evaluation reports in our sample offers recommendations that follow from the analysis and conclusions and are both practical and directed at specific groups of actors (see Table 14). However, as the conclusions, in a number of cases, are based on insufficient evidence, as described above, we need to question the reliability of the recommendations. It must also be noted that almost one third of the evaluations fail to provide satisfactory recommendations.

Recommendations in a mid-term project evaluation will look quite different to those in an evaluation of an intervention that is close to its end, and recommendations provided by a formative evaluation will be quite different to those of a summative evaluation. The evaluation of Sida's activities in the field of culture and media (04/38), for example, has diverse, abstract and long-term recommendations. A programme evaluation has very concrete recommendations for clearly specified actors and for immediate action. What is useful in one evaluation could be out of place or incomprehensible in another. The recommendations must be developed according to the nature and purpose of the evaluation, the project cycle and the kind of decision-makers that are being addressed. It must be clearly stated which value judgements the recommendations are based on.

Having assessed the recommendations according to the three main criteria (above), and acknowledging that recommendations must be seen in their context, let us turn to the evaluations with which there are problems. What is wrong and why? The issues can be summarised as follows:

- In some evaluation reports, the value judgement or judgements underlying the recommendations are not explicitly stated or, if they are tacit, not easily grasped and understood by the reader. This point is important because recommendations on how somebody should act cannot be drawn from observations alone but must be supplemented with value judgments of the following type: given that we value this aid effort and that it should continue, our observations on how it actually works suggest that it should be improved in the A, B and C way.
- Some recommendations are beyond the control of the intended users. There is no point in suggesting actions that are outside the mandate or beyond the resources of those who are to respond.
- The evaluators may need to distinguish between different types of recommendations according to whether the underlying value judgement is discontinuation, long-term continuation or short-term amelioration, and at whom the recommendations are directed.
- Recommendations should consider the costs and benefits of making the suggested changes, including the risks they involve – particularly if major changes are recommended.
- Evaluators need to be careful and prudent in the way they express their recommendations. The choice of words is important. Powerful recommendations can be diluted by an overly meek style, while particularly sensitive recommendations might be dismissed because of an overly assertive style.
- There is a practical limit to the number of recommendations that should be suggested in an evaluation. Absorptive capacity puts a limit on the number that can be digested and acted on, though this will of course vary. Some 3 to 6 highly strategic recommendations followed by some 10 to 20 more operational suggestions would probably be the limit in many cases.

We present what we consider to be reasonably good examples of recommendations in Box 12. These recommendations are made at a point when the programme is approaching its end, but the authors go beyond the programme and outline spheres of cooperation for a longer-lasting relationship. This reflects the fact that Sida frequently asks evaluators to advise on future initiatives.

Box 12. Good practice: Policy recommendations beyond the programme

Twinning Cooperation with the Baltic States Concerning Prisons and Probations

9.2 The future

As we understand it the Sida funding of the twinning cooperation will come to an end when the Baltic States accede the European Union, that is on 1st of May 2004. We find it important that international cooperation can continue at a minimum budgetary level and that also other measures by the Baltic institutions are taken in order to sustain and develop the activities and results of the twinning programmes. Hence, there is need for a strategy for the future. We think that such strategy should be elaborated by the prison administrations themselves in the Baltic States and Sweden, but we will offer some suggestions. The Baltic States should

- strengthen the sustainability of ongoing activities and results. One way of doing so would be to improve the dissemination of knowledge and experience of the twinning cooperation (spread of "best practice") within the whole system in each Baltic country. The staff training centres should play a role in this area of work;
- appropriate budgetary means necessary to finance some activities of external cooperation, for instance for travel to other countries that is necessary to maintain an international cooperation;
- reinforce the capacity to manage aid resources from EU, that is the administrative capacity to design and implement EU-funded projects of use to the prison and probation system, including funding of non-governmental organizations that carry out work in this area;
- involve non-governmental organizations complementary to the State. Such organizations are well equipped to deal with pre-release preparation, aid to newly released persons (work training, studies, social contacts, food etc) and other tasks. They can perform some of these tasks better than State institutions and may also have the possibility to raise money in addition to the State budget;
- develop tools for cooperation that work on a low-budget basis. Internet communication and e- learning are tools that could be used in efforts that are joint for all three Baltic States; and
- prepare for a transition to a regular international cooperation concerning prison and probation.

9.3 2003

We recommend Sida funding at the present financial level during the next year (2003). A reason to continue the cooperation is not only that it has had good results but also that the Progress Reports of the EU Commission point at the justice Sector – for instance the magnitude of pre-trial detainees – as one of the weakest of the Baltic candidate countries. In addition, there are several twinning arrangements of recent date (Maardu, Tartu and Lukiskes), which must have a chance to develop.

With regard to the content of the cooperation in 2003 there are many needs to meet and, hence, many areas of activities that deserve attention. Examples are prevention and combat of drugs; prevention and combat of HIV; probation and other alternatives to imprisonment; prison management; and material. As has been the case in the past, every twinning arrangement should be fairly free to determine the content of its own cooperation. This decentralised way of decision-making is a way to ensure a high degree of relevance of the activities. But there could also be some elements of

cooperation at the policy level that could be continued in the future. A possible topic could be attitudes of the general public towards prisoners and ex-prisoners.

In addition to the continuation of such activities we recommend that the elaboration and implementation of the strategy for the future starts as soon as possible. We find it reasonable that part of the funding available for 2003 be used for the kind of strategic activities that were mentioned above (9.2). In this way the last period of Sida funding will be used to lay the basis for continued cooperation on a low budget basis.

9.4 After 2003

As already stressed, we find it important that some kind of cooperation can continue also after the date of the Baltic States' accession to the European Union. Otherwise much of what has been created may be lost. It should be kept in mind that several activities do not require much funding. Examples are legislative work to create alternatives to imprisonment, dissemination of best practice through out the prison system of each of the countries involved, methodological development including increased awareness of gender aspects and supply of used material. Also some on-job training of Baltic staff in Sweden would be appreciated. A continuation is, however, conditioned on the strategy suggested above.

Such a strategy may pave the way for the transition to a regular, non-subsidized cooperation with other countries and also between the Baltic States themselves. We want to point out that these States must be prepared to make modest contributions of their own for the regular international cooperation. Such contributions may have considerable impact since they make possible informal contacts, visits, e-mail counselling and e-learning, transport of used material etc. It should be a natural thing in the future for Baltic prisons to have exchange with neighbouring countries. Such activities, even though sporadic, are already going on, for instance with Poland and Russia. In addition, it could be beneficial for SPPA to include prison management from the Baltic States in future international cooperation with other States in eastern Europe.

Whilst many Swedish authorities have a self-interest in cooperating with their Baltic counterparts, e.g. police and customs, the interests of Kriminalvårdsstyrelsen (SPPA) are less pronounced but, nonetheless, we find it justified for SPPA to allocate some budgetary means for a regular cooperation in the future, i.e. allowing for a continuation of some of the activities that are now funded by Sida.

Source: SE 03/11. Development Cooperation between Sweden and the Baltic States in the Field of Prison and Probation

An evaluation could be far more direct in its recommendations than the one quoted above. Many project evaluations are expected to deliver inputs to ongoing activities, and direct, specific recommendations might therefore be expected. Box 16 contains an excerpt from an evaluation of a regional training programme in network maintenance. The evaluator studied a programme in Sri Lanka and made recommendations on how it should be modified in the future. The text in the box shows that the evaluation has a number of suggestions for the future design of the programme. It is practical and the course administrators should be able to develop a new programme with these observations in mind.

A recommendation like this is directed at the course administrators, however, and may be very useful to them. Naturally, for Sida, the first question is whether or not to finance the programme. As the recommendations for change are quite significant, it may be assumed that there were several problems with the programme as it was – too theoretical, insufficient equipment, too expensive, etc. Do poor results imply that the project should be closed, or that it should be modified and extended to make sure the objectives are met? An evaluation can give an opinion or make an explicit recommendation, but both of these activities require value judgements in addition to evidence of performance and outcome effects.

Box 13. Good practice: Practical and concrete project recommendations

As stated earlier most participants expressed concerns about the learning environment, in particular the state of the labs, computers and lack of hands-on, practical experience. Therefore, it is recommended that future programmes place a greater emphasis on practical exercises. To facilitate this, it would be desirable to have more networking equipment available to work on in labs and more up-to-date workstations. Moreover, the D[esign,] I[nstallation,] A[dministration] and M[aintenance of] N[etwork Systems] programme needs to be refined to avoid duplication and overlap and build in more practical training. Specific, technical recommendations include the following:

1. The course should include a component with at least DNS, Mail Transport agent, Network File Service, web server application and Authentication and Authorization services.
2. The course should be more practical than theoretical.
3. The course should include network troubleshooting sessions in which network traffic analysers are used as tool to monitor and detect errors in the network.
4. Laboratories need additional equipment in order to provide the required environment for practical sessions.

Source: SE 04/18. The Regional Training Programme in Design, Installation, Administration and Maintenance of Network Systems (DIAMN)

It takes time and good judgement to write high-quality recommendations. It is quite clear that in the evaluation reports we rated highly, the authors spent a lot of time thinking about the recommendations. They are well structured and carefully worded, anchored in conclusions and supported by data. Where the recommendations are poor, it may be that time was running out and a few bullet points were thrown in to satisfy the TOR. This is the impression conveyed by the six reports with a very low rating.

5.4 Lessons Learned

Sida's evaluation policy places a strong emphasis on learning. It is one of the two overarching reasons for evaluations being commissioned in the first place. Evaluations are meant to contribute to long- and short-term learning, and to learning at different levels. Sida is not alone in this. Most other development cooperation agencies also expect evaluations to contribute to learning.

Learning does not occur automatically however. Non-utilisation and underutilisation of evaluation is a constantly recurring lament in the literature. Decision-makers commission and fund evaluations but care little, if at all, about the resulting final reports (Vedung 2000:265 ff.). In many instances, there is a weak correlation between evaluation and learning. One study (Forss, Samset and Hauglin, 1992) found that evaluation ranked no higher than 17 on a list of instruments that contribute to organisational learning (with a possible ranking between 1 as the best and 19 as the worst). Yet other studies have found that learning can occur quickly with the help of evaluation when agencies work together; there is a sense of urgency, and other organisational mechanisms are supportive, but these three conditions are often not in place (Forss, Cracknell and Strömquist, 1997).

Evaluation may lead to learning and improvement within the framework of the intervention and the activities that are evaluated. It may also promote learning in a broader context of other interventions. For this to occur, the sections in the evaluation reports on lessons learned may be instrumental.

Sida defines "lessons learned" as: *"[g]eneralizations based on evaluation experiences with projects, programs, or policies that abstract from the specific circumstances to broader situations. Frequently, lessons highlight strengths or weaknesses in preparation, design, and implementation that affect performance, outcome, and impact."* (Sida Evaluation Manual 2007, p. 112) Lessons learned should thus go a bit beyond the actual project, programme or policy setting, and be of interest to people other than the immediate stakeholders of the evaluation.

There is a footnote accompanying the above quote that clarifies the "lessons learned" further: *"As the term is understood in this manual, the degree of generalisation of a lesson varies from case to case. As the conditions for development cooperation vary, illuminating attempts at generalisation are often restricted to a particular type of context or mode of intervention."* (Sida 2007, p. 112)

Our interpretation of this is that a reasonable way of promoting learning beyond the actual project, programme or policy would be to regard the intervention under evaluation as an example of something more general. This more general theme might be, for instance, a type of intervention similar to the specific project, programme, policy, theme, etc. In order to succeed, however, the evaluators should provide insight into what this more general entity might be.

So, what is the quality like of the lessons learned section in the Sida evaluations? In general, quite poor. No more than five were rated as "quite good" or "excellent" on their presentations of the lessons learned. Many did not have such section at all. Our findings indicate that the lessons learned section is the aspect of evaluation quality in most need of improvement.

Why is there not a lessons learned section in every evaluation report? A glance at some examples of poorly presented lessons learned sections may provide some answers to this question. The evaluation of a project to support local radio stations in Vietnam (04/35) did indeed try to specify "lessons learned" in a particular section, and that in itself is commendable. The quote illustrates what is put forward as "lessons learned" (p. 17 ff.):

> *"Development of human resources is the most meaningful lesson learnt,* which [w]ill be useful for improvement and upgrade of radio programme quality, not only in live programmes but also in any type of radio production. All the selected provincial radio stations have formed up core groups for live broadcasting production. These groups include competent radio staff that can organise and produce live and feature programmes at high quality. In particular, local pioneer broadcasters who have ever been trained in Sweden have made a great progress in methods and attitude of working. After absorbing advance knowledge and skills of modern radio broadcasting, they have become the missionaries who can inspire transfer to their colleagues the knowledge, skills, and their experience.
>
> *"Preparation for and selection of appropriate local radio stations to participat[e] in the project implementation* are also good pra[c]tices that should be shared. Beneficiaries of the pipelined Radio Capacity Strengthening Project at Grass Roots Level will be district and commune stations, which are in disadvantageous positions due to the limited resources compared to the provincial stations. Thus, ensuring availability of resources before implementing project will be an important condition to the success of the project."

In 2004 it was hardly news to the development community that human resource development is important to upgrading radio programme quality. Nor would it have been surprising that time has to be spent selecting and preparing the organisational units to be included in a project such as the one evaluated. However, it might be relevant and important to spell such things out in the local context, and the project personnel there, who might not have been involved in development cooperation previously, had perhaps not thought of such issues before. Although the lessons may have had some relevance in the Vietnamese environment, they would certainly not have been new to either Sida or the Vietnamese authorities responsible for cooperation.

Does this mean that the evaluators should not have documented the lessons learned there? To answer this, the potential audience of the report needs to be considered. The report is in English and is published in Sida's series of evaluation reports. Most of those who read reports in this series are quite familiar with development cooperation. Some of the lessons learned are probably better delivered to local audiences in other ways.

An evaluation of Nine Power Supervision and Control Systems Projects (03/25) presented a set of lessons learned in short and to-the-point statements. We quote three of the six lessons here (they are all quite similar):

> "The donor agency must assume responsibility for managing the preparatory work in such a manner that the development objectives, as opposed to commercial ambitions, become the defining parameters, for project scope and cost…
>
> "It is not reasonable to expect consultants to seriously question the viability of projects that may constitute part of their future market…
>
> "Competent and resource-rich suppliers such as ABB need to be balanced by interests that promote cost efficiency and competitive pricing. This is of particular importance during the project initiation and preparation phase. The record suggests that such a balance has not been achieved in many of the projects under review." (p. 9)

What the authors have to say is indeed interesting and relevant to the project assessment, but are these lessons learned? The discussion on the pricing of supplies under various credit schemes is more than 20 years old and the effects of creating a protected market with combinations of grant aid and credit schemes are well known. The implications concerning what Sida should do and how the agency should undertake preparatory work is not new – but that in itself does not mean that Sida cannot be criticised for shortcomings. The report would perhaps have broken new ground if it had explored why the problems still exist when many earlier evaluations and other studies have identified them and come up with similar recommendations.

Both these examples point to the problem that is actually foreseen in Sida's evaluation manual, namely that the degree of generalisation varies from case to case. What is a lesson to some might be well known to others. A person with many years' experience of development cooperation might not identify any lessons learned at all, while a newcomer with fresh eyes might find lots to learn from. Perhaps it is better to take the chance and present one lesson too many than to be too selective and present too few? There is a risk, however, that many readers will simply skip the lessons learned sections if they constantly fail to yield fresh insights.

When writing a report, it is necessary to make careful distinctions between the three concepts we have considered in this chapter: conclusions, recommendations and lessons learned. As in the two examples above, what are presented as lessons learned are often actually a mixture of conclusions and recommendations. A useful way to handle the formulation of lessons learned is to structure the whole process of inquiry to develop them. We have pointed out the importance of having good evaluation questions – if the questions are relevant it is more likely that the answers will contain lessons to be learned. One commendable way to develop lessons learned is to construct or regard the evaluated intervention as an example of something more general.

Another way to develop lessons learned might be to formulate hypotheses in advance. Evaluators could set down their expectations for their inquiry in the form of up to a dozen hypotheses concerning what they will find. If they are then surprised, find something else or are able to strongly confirm their expectations, the way towards presenting interesting lessons learned is easier.

Furthermore, when authors formulate their lessons, it might be useful to use a "quality test" by inserting a negative into the sentence. If anybody can credibly claim that the negated statement should be acted on, then original sentence can be a relevant lesson learned. If the negative sentence is plainly silly, then the statement may be self-evident. For example: *"Development of human resources…will be useful for improvement and upgrade of radio programme quality, not only in live programmes but also in any type of radio production."* It is doubtful that anyone would claim human resource development to be useless, hence the statement is pointless as a "lesson learned".

What do lessons learned look like when they are well formulated? Box 14 contains an example of good practice of presenting lessons learned. First, the evaluation of Sida's work with culture and media goes straight to the point to assess Sida's policy, which in itself is a rather unusual approach in policy evaluation. Most such evaluations deal with the implementation and effects of a policy, not with its substance. Secondly, the evaluation discusses the policy approach in relation to its context and makes suggestions for further thought.

For those who were working on the strategic development of country programmes, and for many of those who were implementing projects and programmes, the ways in which this evaluation connects the subject to poverty issues must have been quite innovative. Here we can see an example of *"generalizations based on evaluation experiences with projects, programs, or policies that abstract from the specific circumstances to broader situations."*

Box 14. Good practice in writing lessons learned

The Policy for Sida's International Development Co-operation in the Field of Culture presents a thoughtful and progressive view on culture and development that is still "cutting edge" in the international context. However, to be fully relevant to the poverty reduction effort, the Policy would need to be updated and separated into a policy for media, and another for culture.

The new culture policy would need to give thought to how culture can contribute to poverty reduction and empowerment without forfeiting the current Policy's strong human rights perspective and clarity of thought and structure. A stronger poverty perspective should not be interpreted as compromising cultural or artistic merit by reducing the arts to only a function in the service of human development.

The new culture policy would also need to address some additional issues. These include how funds allotted for culture can impact on promoting peace and preventing conflict, support cultural industries and strengthen intellectual property rights. Furthermore, in the light of the Human Development Report 2004 on cultural liberty, a new policy would appropriately elucidate this concept.

Sida's Policy has had the ambition of "making the cultural perspective visible in all development co-operation", which suggests a "mainstreaming" approach. Since culture is a complex concept and naturally varies considerably from country to country, mainstreaming is difficult and can easily be misinterpreted in a way that leads to cultural determinist positions. Thus, at the project level, it would instead be more appropriate to address culture using a rights perspective that focuses on cultural liberty, freedom of expression and freedom of information. Nevertheless, it would be highly relevant to systematically include the roles of culture and media in poverty reduction efforts as standard areas of analysis in the country strategy process.

Sida's culture support portfolio as a whole is highly relevant to the Policy's overall goal of "creating opportunities for cultural diversity, creative activities and sustainable development based on human rights". Human rights perspectives are more prominent in some forms of support than others but, more often than not, the spirit of the human rights framework permeates the support. This includes, in particular, freedom of expression, participation of disadvantaged groups and democratic work processes. The support is also generally coherent with the goal-areas specified in the Policy.

Source: SE 04/38. Sida's Work with Culture and Media, p. 109 ff.

To sum up, even if evaluations are supposed to contribute to learning it might not be possible to generate lessons learned in every case. At times there are severe limitations on the amount of time dedicated to an evaluation and the opportunities for collecting data. Moreover, evaluations are produced and consumed in a context where the majority of readers already has substantial knowledge of the subject or specific activity being evaluated. To produce new lessons for this audience requires not only skills in evaluation but maybe also a better understanding of the specific subject than most evaluators have.

6 Conclusion

6.1 Revisiting the Quality Questions

Let us now return to the questions we formulated in the introduction. Has the study produced any surprises or has it more or less confirmed what we expected all along?

- What information do Sida's evaluations provide, which questions do they answer (q. 1–3, 6)?
- Is the information reliable, can Sida's evaluations be trusted (q. 4–5)?
- Do Sida evaluations support decisions and learning (q. 7–8)?

Question 1. Do Sida evaluations adequately address the evaluation questions formulated by Sida in the TOR?

Most of the evaluations in our sample addressed the questions raised in the TOR, although not necessarily providing satisfactory answers (cf. below). Only six were less than adequate in terms of coverage and none was deemed to have significant shortcomings. As evaluation teams usually present draft reports to Sida, and are asked to make adjustments where necessary, it is not surprising that the end product corresponds fairly well to the TOR.

On the other hand, the TOR were not always clearly formulated and well focused. Our overall assessment of the TOR for the evaluations examined in this study is that they were not very good.

Question 2. Do Sida evaluations provide valid and reliable information on efficiency, effectiveness, impact, relevance and sustainability?

About two thirds of the evaluations contain a satisfactory analysis of effectiveness, sustainability and relevance, but fewer than half contain a satisfactory analysis of impact and only one in five delivers a satisfactory discussion on efficiency. The benchmark "satisfactory" means that the evaluation makes a statement that seems plausible (but would benefit from further elaboration). Thus, for an evaluation the "satisfactory" mark is only a minimal requirement. For certain purposes this level might be sufficient, but in many cases a higher quality of analysis would be desirable. The bottom line is that while the majority of the reports in the sample (74%) were found to address the questions in the TOR, between 30% and 80% of Sida's evaluations failed to deliver plausible statements for each of the five criteria:

- Most of the evaluations cover effectiveness appropriately (62%) although often in the sense of goal-achievement at the output or near outcome stages. Many evaluations that draw conclusions regarding intervention effectiveness did not give the issue of attribution sufficient consideration, i.e. they did not show any empirical evidence of the *intervention* having an influence.
- Impact studies are less common (47%), if we take "impact" to mean the effects of the intervention itself. The issue of causality is the problem of demonstrating that certain outcomes are the result of specific interventions. What effects have occurred as a result of the intervention? Causal analysis should be an integral part of effectiveness as well as impact assessment. Too often, the outcome objectives to be assessed are broad, long-term and of a multiple nature (see q. 2). In many cases the evaluations are designed in a way that makes it difficult to assess the actual impact of an intervention (see q. 3).
- Efficiency is considered to an insufficient extent in most evaluations: only 21% of the evaluations in the sample succeed in this task. Financial analysis is a weak area in most reports and the cost of interventions is rarely analysed and compared to outcomes or impacts – not even at a general level. Questions about the extent to which more and better outcome effects might be achieved with similar or fewer resources using alternative interventions are rarely addressed. The "value for money" perspective should be part of most evaluations however. An overall assessment would be sufficient in many cases. Conclusions are all too often presented without empirical data to support them.
- Not many evaluations apply the sustainability criterion well, and five reports do not discuss it at all. Twenty out of the thirty-four evaluations in the sample are rated as satisfactory. In many cases the analysis would have been more useful if the concept of sustainability had been more clearly defined (e.g. differentiating between organisational and financial sustainability) and a more systematic approach to the assessment of sustainability had been taken. The reports present mostly subjective impressions.
- The assessment of relevance of interventions is found to be more accurate and adequate with two thirds of the sample considered satisfactory. As a minimum, evaluations should discuss programme relevance in relation to needs and consider which alternative and more relevant interventions would be possible. In most cases, however, relevance is assessed in relation to Sida's and the respective partner country's policies. A systematic discussion of relevance with respect to the needs and priorities of the target group is currently lacking.

Question 3. Do Sida evaluations contain a clear and consistent analysis of attribution and explain how and why the interventions contributed to results?

Very few evaluations contain a satisfactory analysis of attribution and causal patterns. Even if they describe impact (which many do not even attempt, see q. 1 above) they follow the logical framework analysis that served as the basis for project planning. The evaluations present and analyse the indicators from the logical framework analysis, but they do not assess the social changes that produce or shape the context in which impact, sustainability and relevance can be assessed.

Question 4. Do Sida evaluations have an appropriate design for impact evaluation?

The choice of design can be considered appropriate in the majority of reports (26 reports are rated in categories 4 to 6), but we were slightly less convinced about the data collection methods (28 reports in categories 3 to 5).

Most evaluations rely on a basic mix of methods: open-ended interviews (94%) and document analysis (91%) are the most commonly used, sometimes combined with ad-hoc observations (35%). Few evaluations use focus group interviews (26%), structured interviews (21%) or surveys (15%). Standardised interviews and structured observations are rare (6% each). Every third evaluation is found lacking in appropriate methods for answering the evaluation questions.

A majority of the reports (85%) gathered data from only a purposive or purely ad hoc sample; the evaluators tended to rely on the literature and documents most easily available. The choice of a sample is very important, but only two evaluation reports contain any discussion on which principles they applied and how the selection affected the findings.

The most common designs are narrative analysis (65%) and case studies (35%). None of the evaluations in the sample used experimental or quasi-experimental designs. For one in five evaluations the design was not considered satisfactory. Impact analysis would in many cases require stronger designs to generate valid and reliable conclusions. We have to conclude that the selection of methods and sources of data collection were not adequate.

Question 5. Is the evaluation process in Sida evaluations well documented and transparent, so that readers can make an independent assessment of validity and reliability?

Fewer than two thirds of the evaluations (56%) contain an adequate section on methods and methodology and even fewer discuss validity and reliability (35%) or limitations of the task (41%). Most of the reports do not include their instruments for data collection or present data to support their conclu-

sions. This means that, in many cases, the reader does not have a chance to make an independent assessment of the evaluation methodology.

There is a need for more and better empirical evidence and systematic use of such information in a majority of the reports. Empirical data provided strong support for the conclusions in only 38% of the reports. The analysis is not sufficiently exhaustive in most of the reports, and it is a weakness that most evaluations tend to use a narrative and descriptive form in the analysis without linking into, or drawing upon, empirical evidence.

For an evaluation report to appear reliable we need to know how the data have been gathered. We did not find sufficient information in the reports on how indicators had been defined.

Question 6. Do Sida evaluations include a valid and reliable analysis of the management of interventions?

An analysis of management aspects is not necessary or relevant to all evaluations. Nonetheless, many of the evaluations include a satisfactory analysis of one or two dimensions of management, while few contain a comprehensive assessment of implementation issues. Fewer than half provide a satisfactory analysis of organisational structures, coordination and networks, and fewer still include a satisfactory analysis of leadership, planning and financial management.

It is striking that leadership and governance issues are often left out or only marginally discussed. A good analysis of leadership and governance issues provides the reader with increased understanding, combining individual and systemic factors. Few evaluations have a specific reference to a gender dimension.

Question 7. Do Sida evaluations provide clear and focused recommendations for specified target groups?

The majority of evaluations have clear and consistent recommendations (82%) that are derived from the analysis and conclusions (71%). As evaluations are often meant to be used for decision-making, it is positive that most of the reports were found to deliver practical recommendations that could be translated into decisions (74%) to clearly specified groups of actors (65%).

As many of the evaluation reports do not have sufficient evidence to support their findings and conclusions (cf. above), however, the quality of the recommendations derived from those must be considered to be questionable.

Question 8. Do Sida evaluations document interesting and useful lessons learned from the interventions that were evaluated?

"Learning" is one of the main purposes of evaluation. The "lessons learned" section in an evaluation report is meant to present new syntheses that are relevant to a wider audience than the immediate stakeholders. Lessons learned are supposed to generalise and extend the findings from the intervention under study, either by considering it as an example of something more general or by connecting it to an ongoing discourse. This requires familiarity with both the international development debate and the discipline or sector under study, and it may not be possible or even necessary in all cases. The degree of generalisation may also vary from case to case.

For all that, it is surprising that only 26% of the evaluation reports contain a section on lessons learned, and it is a cause for concern that the sections that were available are so weak. Only four reports were found to make strong contributions to the understanding and knowledge of development cooperation.

Sida evaluations are diverse: most are good in some respects and less good in others. Some of the authors are quite skilled at building arguments and using their empirical data to support conclusions and recommendations, and others are good at working with figures and tables to illustrate an issue and facilitate understanding and learning. Some evaluation reports have a relevant and reliable analysis but not much information on impact or sustainability and some have a good analysis of implementation but little to say on achievements.

Our conclusion is that there is definitely reason for concern regarding the quality of Sida's evaluations: a majority of the reports were rated at below adequate performance on presenting empirical evidence, justifying methodological choices, arriving at conclusions regarding impact and effectiveness, and documenting lessons learned. Even though the majority of the reports was satisfactory in most respects, most were far from excellent, and there were not many that would be classified as very good on the majority of quality attributes. As the average cost of the evaluations was 780,000 SEK, this represents a significant waste of resources. There are therefore good reasons to try to improve evaluation quality.

6.2 Why are there Quality Problems with Evaluations?

There is a limit to how much we can say about possible improvement on the basis of evidence from this report. While we conclude that evaluations need to be improved, we are less certain about what to recommend and which initiatives will be most important and effective. The following discussion is not based on evidence from our analyses, but is a more open and tentative

synthesis of observations and other people's findings. It is meant as an epilogue to the report and as an introduction to future research on the subject of quality.

This study has helped us to understand what the quality problems are that pertain to evaluation reports, but not what causes them. We have analysed the relationship between costs and quality in the reports as well as team composition (cf. Annex 2), but we have not looked at the processes of selecting consultants or interviewed evaluators or those who commissioned the reports and were the end-users[8]. We have not looked at how and why specific evaluations are proposed and carried out or at the whole process of preparing TOR. We also lack information on the extent to which the evaluation reports were useful and how they were actually used. For many evaluators, usefulness is not a sufficient value criterion: evaluation findings should be useful, but the decisive quality criterion is their actual use. Others are satisfied with less, and argue that usefulness is enough and that actual use is the practitioners' rather than the evaluator's responsibility. There is no systematic analysis of all the possible causes of the weaknesses identified in our report. Further research may provide additional insights into other aspects of quality.

Explanations for inadequate evaluations can easily be found. The threat to evaluation quality can be caused by pervasive, bad practice by individual evaluators. Unlike other professions, evaluation does not have an accreditation system. Anyone can call himself or herself an evaluator and bid for evaluation contracts. Purchasers of evaluations may also lack the expertise to distinguish professional evaluators from well-intentioned amateurs or charlatans. Those who commission evaluations may lack the skills to determine whether or not evaluation products constitute good work. This immediately make evaluators and Sida staff easy targets for criticism.

We believe, however, that it is all too easy to put the entire blame for low evaluation quality on just the evaluators or individual commissioning desk officers, as they do play a significant role in all evaluation processes. There is obviously a broad range of factors determining evaluation quality at all levels of the evaluation system – from the individual evaluator to Sida's evaluation system, as well as weak external demand for high-quality evaluative information. Some possible explanations for the quality problem might be:

- problems in the way evaluations are initiated and the formulation of TOR;
- weak capacity among Sida desk officers to provide technical support to the evaluation process;

[8] Sida has commissioned two reports on these subjects: *Using the Evaluation Tool – A Survey of Conventional Wisdom and Common Practice at Sida* by Jerker Carlsson, Kim Forss, Karin Metell, Lisa Segnestam and Tove Stromberg, published in Sida Studies in Evaluation 97/1; and *Are Evaluations Useful? – Cases from Swedish Development Cooperation* by Jerker Carlsson, Maria Eriksson-Baaz, Ann Marie Fallenius and Eva Lövgren, published in Sida Studies in Evaluation 99/1.

- a limited number of qualified consultants, and missing skills and capacities among the evaluators with which Sida works, which might reflect a lack of competition and little variety among consultants;
- bias, as many of those who evaluate also plan and implement interventions;
- insufficient professional development in the field of evaluation in Sweden – few courses and other training opportunities;
- poor incentives to carry out good monitoring and evaluation – both within Sida and the Swedish embassies, and for evaluators;
- a weak quality assurance systems at Sida;
- a low level of genuine demand for evaluations by Sida and the Swedish embassies.

In other words, there are several possible explanations at various levels. For reasons of simplicity, we suggest grouping them into three levels:

1. the evaluation report and the evaluation process
2. the evaluation system at Sida and the management of evaluations
3. the external demand for and utilisation of evaluations

In order to improve the quality of evaluations, all three levels need to be addressed and the quality assurance approaches adapted to each level. It is easy to suggest practical and immediate solutions on the first two levels, but difficult to change what is happening in the context of an evaluation, demand and utilisation.

6.3 How can the Quality of Evaluations be Improved?

Sida's evaluation system is well established, but the quality assurance mechanisms are still embryonic. The same seems to be true of several other agencies. There is clearly growing concern about the quality of evaluations, though little has been done about it in terms of concrete analyses of evaluative information – and even less about finding the most effective approaches to quality assurance and improvement.

This report has presented a multi-faceted picture of the quality problem but it does not have a straightforward recommendation on the approach to take to improve the quality of evaluation. There are quality issues at different levels, and multiple strategies are required to improve and strengthen quality. We suggest that future work and efforts to improve evaluation quality be concentrated as follows:

- efforts to improve the quality of individual reports produced by external evaluators;

- efforts to assure the quality of the evaluation system; and
- efforts to influence the demand for and utilisation of evaluations.

Each level will require different approaches.

6.3.1 Level one: Improving the reports

The first and most basic level comprises the inherent qualities of the evaluations, which have been the focus of this report. A system for quality improvement and/or assurance to detect and address weaknesses in design, implementation and utilisation is required at Sida. A set of guidelines and standards to guide quality assessment is also required.

Setting guidelines and standards is a common way of enhancing quality. Guidelines and standards are developed and institutionalised by professional evaluation associations such as the European Evaluation Society, the American and Canadian evaluation societies, etc. Central agencies responsible for overseeing evaluations and legislative audit offices in the public sector also develop and adopt guidelines and standards.

We see more potential in formative approaches however. Advisory committees can be used during the conduct of evaluations to enhance their quality, while design issues need to be resolved during the inception phase, by Sida as well as between Sida and the consultants. For the implementation process, a formative system for evaluation quality assurance could be set up using either internal evaluators, line managers at Sida, external evaluators or experts in the fields covered in the reports. The quality assessments would then take place during the evaluation process – assessments of interim and draft reports – in order to produce ongoing feedback and improvement. The final report could also be assessed in order to produce feedback for the evaluators and Sida. Such strategies are already being used, to some extent, as reference groups are often appointed for evaluations.

Internal data quality control practices can also be applied. Such formative tasks can be carried out by internal Sida staff or external personnel, but it is important that they have the appropriate skills and background. Quality often depends on details, and experienced professionals are therefore needed to detect the "killer" details. To ensure follow-up and utilisation of advice, the quality assurance team may also be allowed to enforce sanctions if its guidance is not followed. The challenge would be to design a system for strengthened self-evaluation and reflexivity during the evaluative process, using internal and external resources.

6.3.2 Level two: Improving the evaluation system

Reports are produced by an evaluation system with a range of attributes that have an impact on evaluation quality. At the next level, the focus is therefore

on the characteristics of the evaluation system at Sida, including the guidelines and procedures for preparing and producing evaluation reports, as well as the evaluation process, managed in the Secretariat for Evaluation (UTV) or any given department.

The aim is to improve and strengthen the quality and credibility of the evaluation system as a whole and build up evaluation capacity among the staff and in Sida's departments. At this level, the important thing is to concentrate on those variables or system properties that have the most direct bearing on evaluation quality, e.g. the preparation of TOR, the expertise of those commissioning evaluations, the clarity and relevance of evaluation guidelines and manuals, etc.

While the first level is concerned with the quality of individual evaluation reports, the attention now moves to the system within which the evaluation process occurs. At this level there is a need to assess and strengthen the system and to develop system-level instruments for quality assurance. Some of the key issues for concern are:

- Integration of evaluation into overall planning and management
- Securing sufficient resources for evaluation – both financial and human
- Mechanisms for quality assurance of evaluation
- Utilisation and communication of evaluation results

It has become increasingly common to examine systems and procedures for enhancing the quality of evaluative information. Some years ago, the Swedish Agency for University Affairs scrutinised the evaluation systems of Swedish universities and institutes of technology instead of directly assessing teaching and research. In many countries, audit offices have reviewed the production of evaluations in their jurisdictions. Some international organisations seek certification through a process such as the one set out by the International Standards Organization (ISO standards). Such certification is believed to provide a level of quality assurance of the organisation's products, including evaluative information.

6.3.3 Level three: Improving the demand for and utilisation of evaluations

Evaluations also respond to external needs and demands – politicians, policy-makers and project managers are all stakeholders and users of evaluations. They require relevant evaluation information at a time and in a way that is in line with their needs. For them quality is linked to inherent characteristics of the evaluation report, but more broadly it is linked to the perceived relevance of the evaluation. It is not enough for an evaluation to meet internal quality criteria pertaining to the report itself. A good evaluation needs to come at the "right" time for the stakeholders and address issues that

are on the agenda and need to be resolved. Here, it is the added value of the evaluation system as a whole that is being considered. Indirectly, it may ensure that the other two levels of quality assurance remain user oriented.

Relevant questions:

- Are accurate, timely and reliable evaluation reports produced?
- Does the information reflect the concerns of the various stakeholders?
- To what extent are evaluations utilised? Do findings feed into policy discussions and decisions? Are recommendations reflected in the planning processes of programmes and projects?
- Do evaluations and the evaluation system contribute to strengthening the demand for further evaluations?

It could be argued that such questions belong to the second level with regard system improvements. In practice they do, but we would like to make an analytical distinction between the two since the third level brings in the external perspective of the users of evaluations. While it is much easier to work on systems for improving individual evaluation reports than to make an impact on the demand for and utilisation of evaluations from a long-term perspective, the latter is probably the more important.

6.4 Direction of Future Studies

In this report we have defined quality, identified aspects of quality and assessed the quality of a number of evaluation reports. It is important to keep the debate on quality going and to engage as many actors as possible. In particular, Sida needs to engage the consultants who have been commissioned and the people who commission them in an ongoing discussion on evaluation quality. The challenge is to find innovative ways to generate, develop and sustain an interest in the quality of evaluation.

At its core, quality is a very practical thing. It can be specified and discussed. Most people have opinions on the quality of evaluations. When people at Sida read an evaluation report, they immediately form an opinion of whether it is good or bad. They might not have a list of criteria such as the one we have developed, but many of our criteria would be part of the common sense approach of readers of evaluation reports. The challenge is to give these "common-sense assessments" depth and significance.

Utility and actual use are two related criteria that normally rank high on a list of what makes evaluation "good" or "bad" (= useless). Sida spends money on evaluation, not only to gather useful information, but also so that the information can be put to practical use. Whether and how evaluative information can be used, however, is closely connected to other quality aspects, such as the accuracy of findings. Against this background of knowledge

about quality, it should be possible to undertake more work on the use of evaluations and other sources of evaluative knowledge in the management of development cooperation. The issues of utility and use, for instance, continue to pose challenges, including what should be meant by "use" and which factors facilitate and limit use. Another issue that deserves further study is when and how use is triggered by the evaluation processes that precede the final reports. The topic of process use requires further research.

References

Bamberger, M. et al (2004) 'Shoestring evaluation: Designing impact evaluations under budget, time, and data constraints'. *The American Journal of Evaluation* Vol 25 (1): 67–85.

Forss, K. and Carlsson, J. (1997) 'The quest for quality – or can evaluation findings be trusted?' *Evaluation* Vol 3(4): 481–501.

Forss, K., Cracknell, B. and Strömquist, N. (1997) Organisational Learning in Development Cooperation: To generate knowledge and put it to use. A study commissioned by the Ministry of Foreign Affairs, Stockholm.

Forss, K. and Uhrwing, M. (2003) Kvalitet i utredningsväsendet – utveckling och tillämpning av en model för att bedöma kvaliteten på kommittéarbeten. Stockholm, Regeringskansliet, Förvaltningsavdelningen.

Joint Committee on Standards (1994) *The Program Evaluation Standards.* Thousand Oaks, CA.: Sage.

OECD/DAC (Organisation for Economic Cooperation and Development/ Development Assistance Committee) (2007) *DAC Evaluation Quality Standards.* www.oecd.org/dac/evaluation

Patton, M.Q. (1997) *Utilization-focused Evaluation: The New Century Text* (3rd ed.). Thousand Oaks, CA: Sage.

Rossi, P.H., Freeman, H.E. and Wright, S.R. (1979) Evaluation: A Systematic Approach. London: Sage.

Samset, K., Forss, K. and Hauglin, O. (1992) Learning from experience – a study of the feedback from the evaluation system in the Norwegian Aid Administration. Oslo, Ministry of Foreign Affairs.

Scriven, M. (1993) *Hard-Won Lessons in Programme Evaluation.* New Directions for Program Evaluation, No 58, San Francisco: Jossey-Bass.

Swedish International Development Cooperation Agency (2007) *Looking Back, Moving Forward. Sida Evaluation Manual.* 2nd revised edition. Stockholm, Sida.

Vedung, E. (1997) *Public Policy and Program Evaluation.* New Brunswick NJ and London: Transaction.

Annex 1 Assessment Format: Indicators of Aspects of Quality in Evaluation Reports

1. Assessment of Methodological Choices

	1	2	3	4	5	6	ND	NR
TOR & Evaluation Questions								
Are the terms of reference clear and focused?								
Does the evaluation interpret and focus the task as defined in the terms of reference?								
Is the basic question clearly stated in a specific section?								
Can the informed reader arrive at an understanding of the basic question?								
Description of methods								
Is there a section that describes the methodological choices fully?								
Is there a discussion of threats to reliability and validity?								
Can the reader make an independent assessment of the evaluation methods?								
Is there a clear statement of limitations to the evaluation?								
Design and data collection methods								
Is the design of the evaluation appropriate, given constraints of budget, timing, preparatory work?								
Is the design explained and assessed?								

Are the data collection methods chosen appropriate to answer the evaluation questions?								
Is there a relevant and adequate selection of sources of data?								
Does the choice of data collection methods suggest that the evaluation will get reliable and valid data?								
Instruments								
Are the instruments for data collection well designed?								
Are indicators appropriate?								
Are benchmarks fair and relevant?								
Are rating scales well designed?								

Comments:

2. The Evaluation's Analysis and Assessment of the Intervention

	1	2	3	4	5	6	ND	NR
Analytical content								
Does the evaluation present empirical material in the report?								
Is the analysis relating to the evaluation questions exhaustive and complete?								
Are findings and conclusions supported by the data?								
Analysis of management								
Is there a trustworthy analysis of leadership and governance?								
Is there a trustworthy analysis of planning?								

Is there a trustworthy analysis of financial management								
Is there a trustworthy analysis of coordination?								
Is there a trustworthy analysis of networks and linkages?								
Is there a trustworthy analysis of organisational structures?								
Analysis of achievements								
Is there an accurate assessment of efficiency?								
Is there an accurate assessment of effectiveness?								
Is there an accurate assessment of impact?								
Is there an accurate assessment of sustainability?								
Is there an accurate assessment of relevance?								
Is there a trustworthy discussion of causal patterns?								

Comments:

3 Conclusions and Recommendations

	1	2	3	4	5	6	ND	NR
Does the evaluation respond to the questions in the terms of reference?								
Are the conclusions in the evaluation clear and consistent?								
Do the recommendations follow from the analysis and conclusions?								
Are the recommendations practical, can they be translated into decisions?								
Are there recommendations for clearly specified groups of actors?								
Are there relevant and for an informed audience interesting lessons learned in a specific section?								
Can an informed reader identify and make sense of lessons learned through the intervention?								
Has the evaluation added to a general understanding of development co-operation?								

Comments:

Key to ratings

6 – excellent

5 – adequate

4 – minimally adequate

3 – not quite adequate

2 – significant problems

1 – very poor (or not done at all)

ND – not done

NR – not requested

na – not applicable

Annex 2 Assessment Results: Rating of Evaluation Reports

1. Assessment of Methodological Choices

Evaluation Report	TOR & Evaluation Questions				Description of Methods			
	Are the terms of reference clear and focused?	Does the evaluation interpret and focus the task as defined in the terms of reference?	Is the basic question clearly stated in a specific section?	Can the informed reader arrive at an understanding of the basic question?	Is there a section that describes the methodological choices fully?	Is there a discussion of threats to reliability and validity?	Can the reader make an independent assessment of the evaluation methods?	Is there a clear statement of limitations to the evaluation?
02/06	5	ND	4	5	6	5	5	4
02/07	4	4	3	4	4	ND	4	ND
02/12	5	3	3	5	2	1	4	2
02/35	4	6	5	6	1	1	1	1
03/01	3	ND	3	4	3	ND	3	ND
03/05	4	5	5	5	5	4	5	5
03/09:1	3	5	ND	6	6	4	5	6
03/11	5	1	5	5	4	1	5	4
03/12	5	2	ND	5	3	4	4	3
03/19	5	5	5	5	5	4	5	5
03/25	5	1	4	4	5	4	4	1
03/27	5	3	ND	4	3	ND	3	3
03/29	2	4	3	4	2	2	4	3

03/35	4	2	1	4	2	1	1	1
03/38	4	ND	5	4	5	ND	4	4
03/41	3	3	5	6	5	ND	5	ND
04/04	5	ND	5	5	4	ND	5	ND
04/10	4	5	5	5	5	4	5	5
04/14	4	2	2	4	4	1	1	2
04/18	3	4	2	4	4	3	6	ND
04/21	4	5	2	4	3	3	3	3
04/22	5	6	6	6	5	4	5	6
04/23	4	5	4	4	3	1	3	4
04/24	4	1	5	3	2	ND	3	ND
04/29	4	3	3	5	4	4	4	3
04/32	4	1	6	5	3	4	2	2
04/33	5	ND	ND	5	ND	ND	3	4
04/35	3	3	2	2	2	ND	1	ND
04/36	5	4	5	5	4	4	4	4
04/38	5	5	ND	4	6	6	6	6
05/04	5	1	1	3	2	1	3	1
05/13	5	4	4	5	4	ND	4	5
05/14	5	5	5	5	2	ND	3	4
05/16	5	4	4	4	5	ND	5	2
Summary	**145**	**102**	**112**	**154**	**123**	**66**	**128**	**93**

Key to ratings, see page 87

Evaluation Report	Design and data collection methods					Instruments			
	Is the design of the evaluation appropriate, given constraints of budget, timing, preparatory work?	Is the design explained and assessed?	Are the data collection methods chosen appropriate to answer the evaluation questions?	Is there a relevant and adequate selection of sources of data?	Does the choice of data collection methods suggest that the evaluation will get reliable and valid data?	Are the instruments for data collection well designed?	Are indicators appropriate?	Are benchmarks fair and relevant?	Are rating scales well designed?
02/06	6	6	5	5	5	NR	5	5	ND
02/07	5	ND	ND	4	4	NR	ND	ND	ND
02/12	3	2	4	3	3	3	na	na	na
02/35	2	1	1	3	1	1	na	na	na
03/01	4	ND	4	4	3	NR	NR	NR	NR
03/05	5	4	5	5	4	ND	4	ND	ND
03/09:1	5	6	6	4	5	5	5	5	5
03/11	3	1	3	2	2	na	na	3	na
03/12	6	4	5	5	5	NR	3	3	3
03/19	4	4	5	4	5	ND	5	ND	ND
03/25	5	3	5	4	4	na	5	4	5
03/27	5	3	5	5	4	ND	5	ND	ND
03/29	3	ND	4	4	4	ND	2	ND	ND
03/35	2	1	2	2	2	na	na	na	na
03/38	5	ND	5	5	5	NR	NR	NR	NR
03/41	5	2	4	4	NR	5	5	NR	NR
04/04	5	3	5	5	5	NR	NR	NR	NR
04/10	5	4	5	5	4	ND	4	ND	ND
04/14	3	1	4	3	3	na	na	na	na
04/18	5	3	5	5	4	3	5	4	2

04/21	4	3	4	5	5	5	4	ND	ND
04/22	5	5	4	4	4	na	na	na	na
04/23	5	3	4	4	3	na	na	na	na
04/24	4	ND	3	3	3	NR	2	ND	ND
04/29	5	2	3	5	4	na	na	na	na
04/32	4	1	3	3	3	na	4	na	3
04/33	4	ND	4	5	3	ND	ND	ND	ND
04/35	ND	ND	2	2	1	NR	NR	ND	ND
04/36	5	4	4	4	4	NR	2	ND	ND
04/38	4	5	5	6	6	5	ND	ND	ND
05/04	3	1	3	3	3	na	2	na	na
05/13	4	ND	4	5	5	2	ND	ND	5
05/14	5	ND	4	4	4	ND	4	ND	ND
05/16	4	ND	2	2	2	5	5	NR	5
Summary	**142**	**72**	**131**	**136**	**122**	**34**	**71**	**24**	**28**

Key to ratings, see page 87

2. The Evaluation's Analysis and Assessment of the Intervention

Evaluation Report	Analytical content			Analysis of management					
	Does the evaluation present empirical material in the report?	Is the analysis relating to the evaluation questions exhaustive and complete?	Are findings and conclusions supported by the data?	Is there a trustworthy analysis of leadership and governance?	Is there a trustworthy analysis of planning?	Is there a trustworthy analysis of financial management?	Is there a trustworthy analysis of coordination?	Is there a trustworthy analysis of networks and linkages?	Is there a trustworthy analysis of organisational structures?
02/06	5	5	5	4	4	ND	6	6	5
02/07	5	4	5	ND	ND	ND	ND	ND	ND
02/12	5	5	4	5	5	3	3	3	4
02/35	2	2	2	1	4	4	1	1	1
03/01	4	3	ND	3	ND	ND	ND	ND	4
03/05	4	3	3	5	4	5	5	5	4
03/09:1	6	6	6	5	5	ND	4	4	4
03/11	5	5	5	5	4	3	5	4	2
03/12	4	5	5	3	4	2	3	4	4
03/19	5	4	5	3	4	ND	4	4	ND
03/25	6	5	5	na	na	na	na	na	na
03/27	5	5	5	4	4	ND	4	4	5
03/29	4	4	4	6	5	5	4	5	5
03/35	3	3	2	3	3	2	2	2	2
03/38	4	4	4	5	3	5	2	ND	5
03/41	5	4	4	ND	ND	ND	ND	ND	ND
04/04	4	5	5	ND	5	5	5	5	6
04/10	4	3	3	5	4	5	5	5	4
04/14	2	5	2	1	2	1	2	1	2

04/18	5	4	5	ND	ND	ND	ND	ND	ND
04/21	4	4	5	4	4	ND	ND	ND	4
04/22	6	5	6	4	4	1	4	5	4
04/23	4	4	4	2	2	2	2	2	2
04/24	4	3	4	2	2	2	2	2	2
04/29	4	4	4	5	6	6	4	5	5
04/32	3	2	3	4	4	4	4	3	3
04/33	3	6	3	ND	ND	ND	5	5	5
04/35	2	1	1	ND	ND	ND	ND	ND	ND
04/36	4	4	4	5	4	4	4	4	4
04/38	4	5	5	NR	5	ND	5	5	5
05/04	4	2	3	ND	ND	ND	4	3	4
05/13	4	4	4	4	4	4	5	4	4
05/14	4	4	4	ND	4	5	4	5	5
05/16	4	3	4	3	4	ND	5	5	4
Summary	**141**	**135**	**133**	**91**	**103**	**68**	**103**	**101**	**108**

Key to ratings, see page 87

Evaluation Report	Analysis of achievements					
	Is there an accurate assessment of efficiency?	Is there an accurate assessment of effectiveness?	Is there an accurate assessment of impact?	Is there an accurate assessment of sustainability?	Is there an accurate assessment of relevance?	Is there a trustworthy discussion of causal patterns?
02/06	ND	5	5	ND	ND	ND
02/07	3	5	4	ND	5	ND
02/12	2	3	3	2	5	3
02/35	1	1	1	1	1	1
03/01	3	4	4	4	5	ND
03/05	3	4	4	4	5	ND
03/09:1	5	?	6	6	6	6
03/11	4	5	2	5	5	1
03/12	5	5	5	5	6	ND
03/19	3	4	4	4	4	ND
03/25	6	5	1	2	4	1
03/27	3	5	4	5	5	ND
03/29	ND	4	5	2	5	ND
03/35	2	2	1	1	1	1
03/38	NR	NR	NR	NR	NR	NR
03/41	ND	4	4	4	4	4
04/04	NR	NR	NR	NR	NR	NR
04/10	3	4	4	4	5	ND
04/14	na	na	na	na	na	na
04/18	3	2	ND	4	ND	ND
04/21	ND	5	5	3	5	ND
04/22	4	5	2	4	4	1
04/23	2	1	2	4	3	1

04/24	ND	3	2	1	2	ND
04/29	2	5	6	5	4	ND
04/32	2	4	1	4	3	1
04/33	ND	5	6	3	5	6
04/35	1	2	2	2	2	ND
04/36	3	4	3	3	4	ND
04/38	ND	5	ND	ND	6	ND
05/04	3	2	ND	2	ND	ND
05/13	4	5	4	3	5	ND
05/14	4	4	4	4	4	ND
05/16	3	3	3	4	4	ND
Summary	**74**	**115**	**97**	**95**	**117**	**26**

Key to ratings, see page 87

3. Conclusions and Recommendations

Evaluation Report	Conclusions and recommendations							
	Does the evaluation respond to the questions in the terms of reference?	Are the conclusions in the evaluation clear and consistent?	Do the recommendations follow from the analysis and conclusions?	Are the recommendations practical, can they be translated into decisions?	Are there recommendations for clearly specified groups of actors?	Are there relevant and for an informed audience interesting lessons learned in a specific section?	Can an informed reader identify and make sense of lessons learned through the intervention?	Has the evaluation added to a general understanding of development cooperation?
02/06	6	ND	ND	NR	NR	5	5	5
02/07	4	5	5	5	4	ND	4	3
02/12	6	6	6	6	6	1	5	4
02/35	3	4	2	4	4	1	2	2
03/01	4	4	5	5	ND	4	3	2
03/05	4	4	4	5	5	3	3	3
03/09:1	6	5	ND	ND	ND	ND	5	5
03/11	5	5	6	6	5	1	5	4
03/12	6	5	5	5	4	5	5	5
03/19	5	5	5	4	4	4	4	3
03/25	5	2	3	2	2	3	3	4
03/27	4	5	5	3	4	ND	3	3
03/29	4	5	5	5	6	ND	5	5
03/35	5	4	2	3	4	1	1	2
03/38	4	4	5	4	3	3	3	3
03/41	5	4	4	4	ND	ND	4	4
04/04	5	5	5	5	5	ND	3	2
04/10	4	4	4	5	5	3	3	3
04/14	2	5	2	4	4	1	3	3

04/18	4	4	5	6	5	ND	4	4
04/21	5	4	5	5	5	5	3	3
04/22	4	4	4	3	3	1	4	3
04/23	4	3	4	4	4	3	3	3
04/24	4	2	3	3	ND	ND	3	2
04/29	6	6	6	6	6	1	6	5
04/32	3	4	3	4	4	4	4	2
04/33	5	5	5	4	3	ND	5	5
04/35	3	2	2	2	ND	2	2	1
04/36	4	5	5	5	5	4	4	4
04/38	6	6	6	6	6	6	6	5
05/04	2	2	4	4	3	1	2	2
05/13	4	5	5	4	5	5	5	4
05/14	5	5	5	5	5	2	2	2
05/16	3	4	3	3	3	4	4	4
Summary	**149**	**142**	**138**	**139**	**122**	**73**	**126**	**114**

Key to ratings, see page 87

Annex 3 Presentation: Structure and Style

1. Introduction

The presentation of an evaluation report may seem less important than its content. Indeed, this seems to be the general opinion of the evaluation teams whose final written reports we have analysed, as this is probably the weakest aspect of their quality. The evaluation teams could have done a better job in this respect. That they did not probably reflects that they did not think it was important. When it came to using scarce resources, priority went to data-gathering and analysis, and content, whereas the effort that could have gone into organising the material into a clear and logical structure and presenting an attractive and reader-friendly report was not considered worthwhile.

By and large, it is probably better to devote time and attention to content rather than to form. Still, it would take very little to improve the reports dramatically, and with that improvement the reports would be more useful. A coherent and well-developed structure certainly facilitates reading and comprehension, and as most decision-makers want to access information swiftly, a structure that allows for selective and quick reading will increase the chances that a report gets used. Appealing metaphors and clear, unambiguous words and sentences help to arouse curiosity about and convey the contents of the report.

The form of the report, its deep logical structure and its linguistic surface, contains different quality aspects that should each be assessed in their own right. It is about choosing a title, developing a structure for the text, devising illustrations, figures and tables to facilitate understanding, using a language that is frank, impartial, varied and interesting to read, and, of course, free from spelling and grammatical mistakes. A report that has these qualities is easier to read and understand, and more usable.

Before turning to the analysis of structure and style, we have to inform the reader that this is probably the most unreliable part of our study. It is more subjective than any of the other parts and one where we often had very different opinions initially. For example, a certain title might sound interesting to one person, whereas another might think that a catchy phrase, a play on words, is merely gloss and something that undermines the serious image that the evaluation should convey. Table 14 provides an overall picture of how we have assessed the structure and style of the presentation of evaluation find-

ings. The number of things that were not done is high when it comes to "helping the reader along", for example with illustrations, figures, tables and diagrams. Few reports are properly referenced.

There are things that look quite good. Most reports are free from grammatical and spelling errors. They are mostly frank and address critical issues head on; only 4 out of the 34 seem to fail in this respect. Few have really good titles, but most titles are satisfactory. The same holds for structure, use of chapter headings, etc. But very few evaluation teams appear to have tried to develop their presentation in innovative ways. What we see is mostly good handicraft, but with little developmental thinking to it, little concern for how to attract and keep a reader, and few tricks of rhetoric to keep the audience interested. In this annex we discuss the details of structure and style and provide some examples of good practice.

Table 14. Overall Assessment of Structure and Style

Questions asked about the Sida final evaluation reports	1	2	3	4	5	6	NA	Total
Does the title of the report reflect the contents of the evaluation and is it well chosen?	0	0	2	15	14	3	0	34
Is there a clear and adequate executive summary?	2	0	3	10	16	3	0	34
Is there a clear and logical structure to the chapters of the report and to the report as a whole?	0	6	4	10	13	1	0	34
Is there a sufficient level of sub-headings to facilitate reading and understanding?	0	4	2	9	18	1	0	34
Are the headings accurate and do they reflect the content?	0	2	3	14	14	1	0	34
Is the text appropriately divided into sections and paragraphs?	0	3	3	11	16	1	0	34
Are illustrations and figures used to facilitate reading and understanding?	17	6	3	6	2	0	0	34
Are tables, boxes and models well designed, clear and accurate?	4	5	6	10	9	0	0	34
Does the report make use of references and is it appropriately referenced?	4	5	6	13	4	2	0	34
Are annexes well structured and readable?	0	4	5	11	10	4	0	34

Is the report free from grammatical and spelling errors?	0	1	2	10	19	2	0	34
Is the language of the report precise, varied and interesting, and free from jargon?	0	5	5	11	10	3	0	34
Is the report frank; does it address issues squarely and head on?	1	0	3	6	21	3	0	34
Is the report written impartially and does it apply different perspectives to the issues considered?	0	1	3	11	15	4	0	34
Have the authors developed the report in creative and innovative ways?	1	4	9	14	6	0	0	34

Key to ratings: 1 – no, very poorly done (or not done at all), 2 – no, significant problems, 3 – not quite adequate, 4 – yes, it can pass, 5 – yes, quite good, 6 – yes excellent, very well done, NA – not applicable, the question was irrelevant to that evaluation, or the issue could not be assessed because of lack of information

Source: The authors' assessment of 34 evaluation reports

2. Titles that are Informative and Generate Interest

As table 14 shows, only two out of our 34 reports chose a title which we believe was not particularly good. The majority of titles were satisfactory, and quite a few were very good. Three titles were deemed outstanding. When we did the assessment, we looked for several subcomponents of what makes a title good. First, it should of course tell the reader what is in the report. It should be *informative*. That is where most of the reports do well; they usually take the name of the activity that is being evaluated and put that as a title. So an evaluation report could be called "Zimbabwe National Network of People Living with HIV/AIDS", or "Partnership Evaluation of Forum Syd 2001–2003". In both cases, the title is informative.

There is a difference between the two examples, the second says it is an evaluation, the first does not. By and large, we think it is better if the title says it is an evaluation. It does not necessarily have to use the word evaluation, it could for example say that it is an assessment or an impact study, a search for results, an analysis of implementation, or something similar. That should set it apart from a project document, a feasibility study, or some other document. However, as all the evaluations are published in the Sida series of evaluation reports it is obvious that they are regarded as evaluations and so it may not be necessary to have the word in the title.

The second criterion of a good title is that it should *sound exciting.* Perhaps it could do so by provoking thoughts, or having some sense of drama. The box below contains the three titles that we ranked as best in our sample; the first one in particular conveys a sense of imminence and a sense of "now", almost like a good newspaper headline. It could attract readers who would not otherwise be immediately interested in the project as such.

Box 19. Examples: Good Titles of Evaluation Reports

"Performing Arts under Siege:
Evaluation of Swedish Support to Performing Arts in Palestine 1996–2003"

"Three Decades of Swedish Support to the Tanzanian Forest Sector: Evaluation of the period 1969–2002"

"Sida's Work with Culture and Media"

Of course people have different opinions on what makes a good title. Most of the evaluation reports in our sample simply give the name of the subject being evaluated but there are a number of other titles like the examples provided in box 19. There are probably different opinions on these. Some are quite obscure and give the reader no clue at all what the report is about, nor even that it is an evaluation. Others are fun and may thus attract some interest from people browsing the databases.

Surprisingly few of the evaluations make use of the possibility of having a *title and a subtitle.* This is a useful device for combining being informative with being a bit more popular and thought-provoking. The title above, "Performing Arts under Siege", is really a very good example. It has both a short and "attention-grabbing" main title and an informative subtitle, saying what is evaluated, and where and when. An evaluation report title could hardly be better, but for one detail: the words performing arts are repeated, first in the main title and then in the subtitle. That could have been avoided.

The two first titles in box 20 could have had subtitles that provided a bit more information on what you might expect to find in the reports. It would hopefully not have scared potential readers away. Whether the catchy phrase appears in the title or in the subtitle does not matter so much, it could work either way. But it is probably more common to have the vague and more thought-provoking statement first, as in the third and fourth examples in box 20, and then to provide the information in the subtitle.

Box 20. Titles where the authors have been innovative

"Completion of a Success Story or an Opportunity Lost?"

"Innovations Wasted or Wastelands Reclaimed?"

"Turning Policy into Practice: Sida's Implementation of the Swedish HIV/AIDS Strategy"

"Donorship, Ownership and Partnership:
Issues Arising from Four Studies of Donor-Recipient Relations"

The list of report titles also suggests some ways that titles should not be written. First, it should be possible to say the title out loud and understand what it means. Second, the title should not contain acronyms that are not commonly known and it should not contain abbreviations. Third, as the titles should be short, they must not be repetitive.

Producing a reasonably good title is relatively easy, and those evaluations that were given ratings of three and four in table 14 could easily have been improved if the authors had thought about it. Still, what is a good title is also a matter of taste, as not everybody would like or be attracted by titles such as those in box 20. The evaluators must choose the title with their specific clients and potential readers in mind, and perhaps the titles that we rated as just satisfactory are exactly what the clients wanted.

Almost all the reports have an executive summary, and there has been a significant improvement in this aspect of quality over the past decade. A 1997 study of Sida's evaluation system (Forss and Carlsson, 1997) found that 25% of a total of 277 reports did not contain an executive summary (or a summary under some other name). In our sample, 2 reports out of 34 do not have a summary, and in one case this is because the evaluation is part of a synthesis study which contains the executive summary. So it is only 1 report out of 34 that did not use the opportunity to provide inputs for decision-makers.

The authors of the evaluations seem to have different views on the *size of a summary;* some make the summaries short and others make them rather long, at more than five pages. There are those who argue that an executive summary should not be longer than one page. Perhaps that would be possible for a 20-page report. But if the substance of the evaluation runs to some 100 pages, it would be very hard to provide a meaningful summary in one page. The size of the summary should be seen in relation to the length and content of the text itself. Our own preference is for summaries that can be read very quickly, and thus a text of more than 2 to 3 pages would most of the time be too much.

The authors also seem to have different views on the *contents of a summary.* Quite a few see the summary as a brief presentation of conclusions and recommendations. Several choose to present all their recommendations in the summary. In our view this is a mistake. A good executive summary should

contain a condensed description of all the major sections of the report: background to the object being evaluated, evaluation purpose and questions, design, methods, data, conclusions and recommendations. It should be a *summary of recommendations,* outlining the strategic thrust. It should not be a detailed description of each and every recommendation and lesson learned.

Many people start by skimming an executive summary, and a large proportion of these may not read much more. But a well-written executive summary could attract some to read more and thus to learn more about what is in the full report. Much as the executive summary should provide accurate and concise information to the reader who is pressed for time, it should also encourage others to read more. It should stimulate interest in the issues addressed in the evaluation – just as the title should. For this reason, it is especially important that executive summaries be well written.

Most of the executive summaries of our 34 reports consist of straight text, divided into paragraphs. It seems as if there is some kind of a taboo on using presentational devices to facilitate reading and understanding such as headings for paragraphs, bullet points, lists, boxes and tables. Perhaps evaluators think that as the executive summary is so short anyway, there is no need to facilitate the reading further. However, precisely because the audience would be readers who are short of time and who should be helped to grasp the subject very quickly, it is essential to use all available means to achieve this. It is not enough that an executive summary is brief and well written; it should also – where possible – use other stylistic devices to communicate to the reader.

There were four executive summaries to choose from among those rated as "excellent", and they share the same characteristic: they are brief, at 2–3 pages. It is interesting to note that there is a relationship between the length of the text and the length of the summary. There was a good 2-page summary of a text of 20 pages, another 2-page summary of a text of 25 pages, a 3-page summary of a text of 60 pages, and a 7-page summary of a text of 102 pages. These could be taken as benchmarks of how to produce an executive summary. The one we choose to present here is a good example also because:

- It uses sub-headings to present its main findings
- It summarises major achievements comprehensively in table form
- There is a summary of both lessons learned and recommendations, but these are set out more fully in the main text
- There is an overview of the evaluated intervention and its logic
- The reasons for the evaluation are laid out, as is the way in which the evaluators worked.

3. Headings, Sub-headings and Paragraphs

A final written evaluation report will be far more useful if it is possible to follow a logical development of an argument. This requires the authors to present their material in a clear sequence, which can be done in many different ways. Sometimes it is useful to start with the overall conclusions regarding the impact of the evaluated activities, and to then work backwards to explain how and why the impact came about. At other times it is more effective to start with a review and analysis of activities, and then conclude with their results. This is a matter of choice and what is best will depend both on the content of the report, the readers and their context, and the way the authors develop the story.

The linear sequence to the presentation is one aspect of structure and the other relates to the conceptual hierarchies. Some aspects of the evaluation report are more general than others, and some subjects should be treated as subcategories. So for example, management is a rather broad category of activities, as is implementation. In most analytical schemes, other activities such as leadership, planning, coordination and network building would be concepts that are subcategories of the more general term "management". It is important that an evaluation report reflect the conceptual hierarchies accurately, as readers can otherwise get confused by the messages and lose their way amongst the data and the findings.

Both the linear sequence of an argument and the conceptual hierarchies used are expressed in text, but highlighted through the choice of headings – at the level of chapters and subchapters. One of the first things a reader sees is the table of contents. Personally we always look at that first, and if the table of contents gives a good overview of the report we are usually keener to read the whole text. We can also choose what sections to focus on, and perhaps which ones to skip. If the report is relatively short, it is very good if the table of contents can fit on one page. If the text is longer, it may be necessary to have more chapters and subchapters and hence a longer table of contents. But if the table of contents is longer than three pages it will not be a helpful tool for the overview.

Box 21 shows an example of a very good report structure. What is it that makes it so good? Let us try to summarise the best features:

- It fits on one page and gives a good overview of quite a complex evaluation topic
- There is a logical progression from description, to assessment, lessons learned, conclusions and recommendations
- There is one major section on management issues and another on results, with a clear distinction between them but also a connection in the concluding chapters
- It uses a sufficient number of headings and sub-headings to give an overview of content

- It uses mainly two heading levels, but in some cases three
- When three levels are used, there is only a small number of headings
- There is symmetry to the presentation and chapters are more or less of equal length, except for the chapters where most of the empirical material is found
- Note that there are between 3 and 6 sub-headings in each chapter, which conveys a sense of balance and serves as an overview of the chapter.

Box 21. Good Example: Clear and Logical Structure of an Evaluation Report

Table of Contents

Source: Sida Evaluation 03/12. Three Decades of Swedish Support to the Tanzanian Forestry Sector: Evaluation of the Period 1969 – 2002.

By and large, this is an example of how a very complex subject, three decades of support to the forestry sector in Tanzania, is made clearly understandable to a broad audience. The structure helps the reader find what is interesting and enables them to see a logical progression of cooperation leading to results and then onwards to the future. Regrettably, it is more common that something relatively simple is made complex by the lack of structure in an evaluation report. So what are the most common mistakes? Well, often they are exactly the opposite of what we saw in box 21.

There are either too few or too many chapters. Some evaluation reports divide the text into 15 chapters or more for a text of 30 to 40 pages, and others use 3 chapters for the same amount of text. In neither case does this simplify or facilitate matters for the reader. In some cases we have seen chapters that are no more than a quarter of a page, followed by chapters of 15 or 20 pages. If there is no more to be said on a subject than what fits on half a page this should not be a chapter in its own right but a section of another chapter.

In Sweden we are used to writing in paragraphs that are distinctly separate, as we do here. It is also possible to have paragraphs that are not separated, but where the next paragraph starts a bit further in. Some say that this provides for smoother reading, whereas the practice we use here automatically gives a staccato rhythm to the text. Be that as it may, both ways can be used and seem to work in evaluation reports.

Box 22. Example: Text Format with Numbered Paragraphs

3.8 This evaluation and the judgements articulated in it are to be understood against this complex Project background. Specifically, the processes involved in the three components – decentralisation, poverty reduction, strategic advice, capacity-building, sustainability, empowerment, civil society, coordination, action learning – are abstract concepts open to a wide variety of interpretations and understanding.

3.9 As experience consistently demonstrates, institutional change is a supremely problematic, long-term process. Decentralisation, with all its claimed potential benefits, is perhaps the most challenging type of institutional change. This is because it inevitably involves changes in the hierarchical location of the two most fundamental elements of any organisation: power and control.

3.10 Decentralisation of authority is 'safe' for the stakeholders while it remains a concept confined to the pages of reports and recommendations. Opposition – overt and covert – inevitably appears when decision-making authority is about to be transferred down the hierarchy, i.e. delegated. ……..

Source: Sida Evaluation 04/33. Swedish Support to Decentralisation Reform in Rwanda.

Some of the reports had British authors, and in their stylistic tradition it is common to number paragraphs. There is an example in the text below. The numbers are used instead of sub-headings, and the reports have page after

page of numbered paragraphs but very little other variation. This definitely interrupts the reading and forces the author to make one statement at a time rather than developing a coherent text. It does have the advantage that it is easy to discuss the text and direct attention to its different parts, but otherwise it is a very cumbersome form of writing.

There should be a sense of symmetry to paragraphs as well. Some authors use paragraphs that consist of no more than one or two lines and this interrupts reading and makes the reports unnecessarily long. Others never end their paragraphs, letting them run over a full page. There are a few who mix excessively long paragraphs with paragraphs with a sentence or two across the same number of lines. Both structures make a text difficult to read and comprehend.

The choice of headings and sub-headings should both denote what the content is and emphasise the structure of the story being told. Needless to say, the choice of words should be made with the readers in mind. There should never be abbreviations in headings, and jargon and technical terms should be avoided. Here evaluation reports face a dilemma as many readers belong to "aid bureaucracy". For us (we are probably part of that environment) it is common to speak in terms of impact, sustainability, and different management terms, and to use these as headings. But for many others, the real world is talked about using other words. Sustainability may be about how to get money, what to sell, how to get and maintain political support, etc.

The headings and sub-headings shown in box 21 are good examples of how a balance can be worked out between the demands of a profession and the common sense of most readers, with an added twist regarding the forestry sector. Who would not be amused by the choice of sub-headings in chapter 3, describing the evolution of forestry cooperation in terms of seeds, sapling, branches and pruning? Quite clever!

4. Using Illustrations, Figures, Tables and Boxes

There is a distinction between on the one hand the use of illustrations and figures, and on the other hand the use of tables and boxes. They all serve to present empirical and theoretical material and to make complex issues more easily understandable; pictures may both illustrate findings and in various ways amuse or stimulate the interest of the reader. It takes more creativity to use pictures and figures, but they make a report more digestible and fun to read.

It is surprising that so few evaluations use illustrations, and some only do so in a rather haphazard way – with no obvious connection to the text or the object being evaluated. There is one exception, and that is the Sida Evaluation 04/32 "Environmental Remediation at Paddock Tailings Area, Gracanica, Kosovo". The authors use photos to illustrate solutions to the environmental problem described in the report.

A figure can be a useful way to illustrate the intervention theory behind an aid effort, to show causal patterns, stakeholder influences, management structures, networks and coordination, etc. A figure can say more than a thousand words – well, almost, at times. Again, very few of the evaluation reports use figures. Those that do gain a lot in clarity of presentation and in helping readers visualise the processes they are describing.

There is a good example in Sida Evaluation 02/35 "Implementation of the 1999–2003 Country Strategy for Swedish Development Cooperation with Vietnam", which is reprinted, with some modifications to fit our format, below. The figure (box 23, following page) can be considered a good example because it:

- makes a complex process simple and understandable
- describes the sequence of events clearly
- clarifies the concepts that are used later in the report and connects them with each other
- shows the progression of time and how long the processes really take.

Tables, boxes and shaded areas are more common than figures in the reports. Authors who present empirical data from questionnaires normally use tables, and it is also common to use tables when financial data are presented and analysed. We found relatively few examples of excellent use of tables. It is an art to design good tables but few have learnt to master it. There are two aspects to consider. The first is concerned with making the table easily readable by having sufficient space in cells, and using clear fonts and varied styles between different forms of entries. The second concerns formalities that should be right; the source of data should be indicated, the size of population and sample should be provided, and the totals should be added up correctly and "no-responses" or drop-outs should be clearly identified.

Box 24 below reproduces two tables from Sida Evaluation 03/35 "Sida's Support to the University Eduardo Mondlane, Mozambique". We feel they are good examples of tables, and could be presented as good practice. Why?

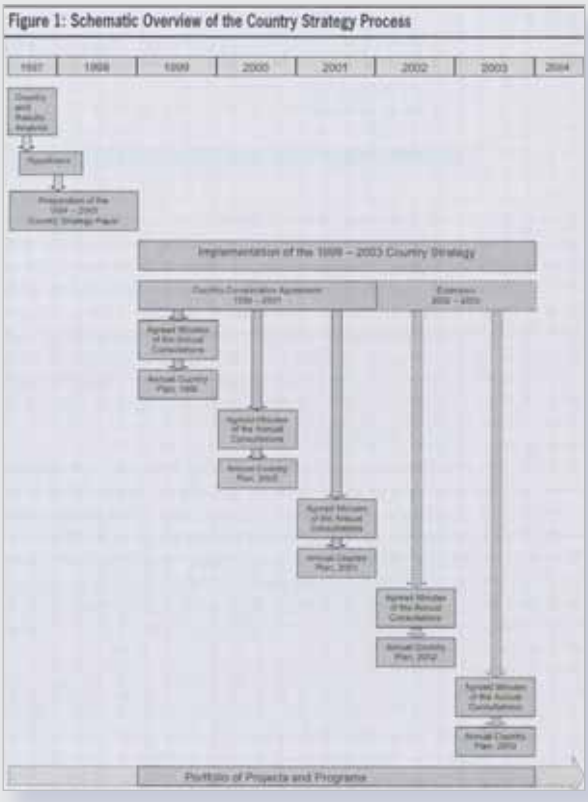
- They have clear titles that are put above the tables so that it is immediately possible to recognise what the table is about.
- The design is not overcrowded with information, but the data are detailed enough to provide an overview of the subject.
- The column headings are in bold type, which identifies the nature of the columns clearly.
- There are rows and columns for the total figures in both tables.
- Table 8 shows how a column can be subdivided into three levels, without confusing the reader. It brings out relevant distinctions, and it is particularly good that it shows how to easily present gender-disaggregated data.

- A table should have information on the data sources. Table 4 has that information but not Table 8 (it is in the text, but should be in direct connection to the table also).

Many of the reports put tables and diagrams in annexes. When the material to be presented is rather extensive this could be a good idea. However, the main purpose of tables and figures is to economise on space and to make complex issues more easily understandable. So, if the table fulfils its purpose, it would be better to have it in the main text. Annexes should be used for parts of the evaluation that are not of immediate and obvious interest, for things of secondary importance, or for things of specialist interest.

There are many guidelines available on how to present data in figures, tables and visual presentations (see for example Torres, Preskill and Piontek (1996) for a handy and practical guide). Some of these (and other) hints are summarised in box 25 below. The last item in the list is probably where most authors go wrong. Much as tables and figures are ways of making things more easily and clearly comprehensible, they are still difficult to do well.

Box 23. Example: Use of a Figure to Simplify Complex Processes



Source: Sida Evaluation 02/35. Implementation of the 1999–2003 Country Strategy for Swedish Development Cooperation with Vietnam, p. 4.

Box 24. Examples: Good Practice – Tables

Table 4 Mozambican publications produced in international journals over 1981–2002.

Subject categories	1981–85	1986–90	1991–95	1996–00	2001–02	1981–2002
Sciences	34	66	89	165	73	427
Social sciences	25	15	31	20	5	95
Arts and Humanities	9	4	6	2	2	23
Total	68	85	126	187	80	546

Source: ISI (22 November 2002)

Table 8 Students and teachers by gender at UEM, at all public universities and at private universities in Mozambique 2000–2001

Gender composition at Mozambican Universities	Students					Teachers				
	Men		Women		Total	Men		Women		Total
	No.	%	No.	%	No.	No.	%	No.	%	
UEM[42]	5 430	74	1 877	26	7 307	585	77	175	23	760
Total public universities	6 735	74	2 241	26	8 976	785	76	245	24	1 030
Total private universities	2 269	53	2 013	47	4 282	331	75	104	25	435

Note: ISI = International Institute for Scientific Information;
UEM = University Eduardo Mondlane

Source: Sida Evaluation 03/35. Sida's Support to the University Eduardo Mondlane, Mozambique, p 29, 53

Tables and figures are economical as tools of communication, but that means they save time and energy for the reader, not necessarily for the author. It is probably more time-consuming to develop a model and a good illustration than it is to just produce the text. But the authors serve a larger public and should take the time to do that well. Constructing a decent table may, even with the help of Microsoft's different tools, take several hours. Figures are even more time-consuming. There is no such thing as a free lunch.

Box 25. Guidelines: How to Use Tables and Figures

1	Think about the essence of the message and the type of presentation that will describe it most accurately and effectively
2	Consider if more than one table or figure is needed to communicate a particular set of data
3	Include captions for all tables and figures
4	Make each table and figure self-explanatory by providing captions, keys, sources
5	Construct the tables and figures first, then write the accompanying text
6	Make tables and figures accessible within a report
7	Do not overuse colour, patterns, lines around cells or fonts
8	Allow sufficient time for developing tables and figures

Source: Torres, Preskill and Piontek (1996) Evaluation Strategies for Communication and Reporting. Sage, London

5. Style

Writing style is a highly individual choice and it is not really fair to assess it; a style of writing that appeals to one person might seem repulsive, arrogant and ironic to another. What is clear and direct for one may be technical and jargon-laden for another. We have tried to be very careful when assessing the texts and to stick to objective criteria that many – if not all – could agree on. These are:

1. that the report should be free from spelling errors
2. that it should be free from grammatical errors
3. that it should be clearly written, that is, the messages should not be confused by long sentences or sentences with several clauses
4. that it should be free from technical jargon
5. that it should be frank, and if there is a need to be critical it should not hide the criticism with statements that belittle or express doubts about the findings
6. that it should be impartial and try to see things from several perspectives.

Most of us who write evaluation reports are not authors by profession. We come to this task as economists, social scientists, engineers, environmental scientists, statisticians, civil servants, agronomists, medical doctors, etc. We may be used to writing, but our specialties lie in substance and methodology rather than form. Furthermore, in development cooperation evaluation most of us do not write in our native tongue.

Given this background it is surprising that the majority of the reports are relatively free from errors and free from jargon. No more than a small minority of five or six reports were found to be less than satisfactory in these respects. Nevertheless there are problems: for example, reports are mostly clear, but as non-native English speakers we have a tendency to write in the indirect form and to use the passive tense. That makes the text less interesting, even if it is not wrong per se.

The most positive aspect of style is that a large majority of the reports are both frank and impartial. These are two very important aspects of the writing quality of an evaluation. To take an example from Sida Evaluation 04/29 "Mozambique State Financial Management Project (SMFP)" (pages 55 and 56):

> "9.9.1. The first half of the evaluation period was about reinstating the existing accounting arrangement and making it effective. We conclude that this was the correct approach. It was non-controversial and saw important improvements.

> 9.9.2. The second period, however, was much more involved. We conclude that SISTAFE 1[State Financial Administration System] was feasible and operable. In our opinion though it was a lot for M[inistry of]P[lanning and]F[inance] to ´swallow´ and would have been better presented as a phased programme over several years. No doubt a sustained implementation programme and training would have been implemented, something for which the project has demonstrated capacity.
>
> 9.9.3. An unwritten question contained within the T[erms]O[f] R[eference]s was whether the project was trying to impose a uniquely Swedish model. In fact SISTAFE 1 was an adaptation of the Portuguese national system and chart of accounts. Where the project may have placed particular emphasis however, was holding up an accrual standard as the goal at which to aim. To the Swedish project this was synonymous with modernisation and consistent with the long term objective (goal) of the project.
>
> 9.9.4. However, no developing country government has yet successfully implemented accrual accounting and it is recognised that the majority are a long way from implementation. Its advantages over reformed cash accounting are realisable only in a relatively sophisticated and performance-oriented environment. In our view,* cash accounting is sufficient and does not preclude modernising improvements…."
>
> * Based on accepted wisdom of the World Bank FM specialists and leading figures in IMF Fiscal Affairs Department and the development community generally.

This is a very significant criticism. It says that the project was not effective; it did not do "the right things". It tried to introduce an accounting system that was too sophisticated for the environment in which it was to operate. And as we are dealing with national accounting systems, it is not a minor thing to be wrong about. The authors say so clearly. Their language is direct and straightforward, but at the same time it is not written provocatively or aggressively, as can sometimes happen. It is a good example of how to express dramatic finales.

The text contains a reference to World Bank and International Monetary Fund specialists and publications to support the comparative statement. This was one of the few evaluations that were properly referenced. It contained footnotes where literature on public accounting, financial management, capacity-building and technical cooperation was quoted. The authors had a good grasp of these subjects and used the literature to provide benchmarks and to support their own hypotheses. Few evaluations worked actively with references, in fact only 6 out of 34. That is rather surprising, as it would make the task itself easier for the evaluators. Working with references would

allow more accurate assessment and would instil confidence among readers by referring to similar examples elsewhere.

To illustrate some style issues, we use the quote below, from Sida Evaluation 05/04 "Regional Training Programme in Environment Journalism and Communication in the Eastern African region". The quoted text is typical of the report: sentences are very long, there are long rows of nouns – nouns that each signify rather complex issues – and the reader is left wondering what did happen. The author is certainly not wrong in substance, nor misleading, but the style of presentation makes the message obscure.

> "The programme environment is part of much larger processes involving and evolving a complex and dynamic environment that includes social, cultural, political, economic, legal, technological, and physical, biological and man-made environments. Each of these environments involving groups and individuals with their own goals, purposes, aspirations, desires, motives and resources to influence the outcome of desired long-term developmental objectives. The outcome and impact on individuals, in the societies at large, and within the region depends on the quantity and quality of these interactions. The eventual possibility to implement and sustain the outcome of this programme depends therefore on an enabling environment within this larger context.
>
> There are needs, wants, considerations, bottlenecks, and challenges to be met in this programme at strategic, policy and implementation levels. There are also needs in practices of management and administration, implementation of activities, monitoring and accountability within the programme, within the funding agency and between the programme and Sida. The context and motivations to the recommendations are provided in the text in respective subsection of this document."

There are many useful guidelines on style, and in box 26 we list some of these. However, most evaluators could easily find such guides if they were interested. The question is: why do the majority of evaluators not spend more time developing their writing style? Probably because they believe it is good enough as it is. And here we do not really say anything else. Most of the evaluations are rated at level 4 or above, that is, they are OK – no more and no less. If time is short and the audience limited, why bother? If people had more time, they would probably give priority to developing the methodological and substantive aspects of the evaluation. The style of presentation, clear sentences that communicate well, would probably rank third or fourth on a list of priorities.

6. Concluding Remarks

Returning to table 14, the overall picture is that the majority of the final evaluation reports are rather mediocre when it comes to structure and style. There are several authors who do not bother to use figures, illustrations and diagrams, and many who have not invested enough time and creativity. We have hinted that the explanation for this probably lies in the fact that those who write the reports are more attentive to substance, and sometimes to methodology, than to presentation.

Nevertheless, evaluation reports should be written to make an impact. It is probably true that many reports are not widely read, and perhaps no amount of effort will make them bestsellers. However, that cannot be taken as an excuse not to try. Some reports, some of the time, are not read because they are so boring, so clumsily written, and with so little in the way of structure, references, illustrations, headings, etc to attract the reader. We should remember that they are all published in Sida's Evaluation Report series, which indicates that there is an ambition to disseminate them to a larger audience.

Box 26. Sources of Reading on Style

Author	Title	Publisher and date	Comments
Joseph M. Williams	Style: Toward Clarity and Grace	University of Chicago Press, 1999	The cover says it is a master teacher's tested program for turning rough drafts and clumsy prose into clear, powerful and effective writing. No boasting, it is true, but the focus is on language only.
Rosalie T. Torres, Hallie S. Preskill, Mary E. Piontek	Evaluation Strategies for Communicating and reporting	Sage Publications, 1996	This book covers all aspects of how to write reports – and present findings in other ways too. It discusses structure, executive summaries, tables and figures, annexes – the works!
Lynne Truss	Eats, Shoots & Leaves: The Zero Tolerance Approach to Punctuation	Profile Books, 2003	Really amusing and a much-needed reminder of how to use commas, semi-colons, exclamation marks, full stops, etc. It is only about language though.
Kingsley Amis	The King's English: a guide to modern usage	HarperCollins, 1997	Classic and a bit conservative perhaps, but essential and accessible. The fact that this is a well-known author gives it an extra advantage. This author can practice what he teaches.

Again, we have to address the questions of whether the criteria mentioned above should apply to all evaluation reports, and whether they should be applied to all reports in the same way. To take the latter question first, there are many ways to structure a report well, many ways to use figures, illustrations, tables, etc. There are also many different writing styles, many ways to communicate well. However, this can all be considered under an overall quality criterion: it is important to have a good structure to the report, the text should be free from errors, and so on. The criteria in themselves are absolute, and it is hardly possible to find reports where these attributes do not define quality.

But, practically and pragmatically speaking, are the quality criteria always equally important? It is vital that a report such as "Sida's Work with Culture and Media" ranks high on these quality criteria – it is meant for a wide audience and should facilitate learning by many inside and outside of Sida. However, a report such as "Environmental Remediation at Paddock Tailings Area, Gracanica, Kosovo" may not be read by more than a handful of people. The latter evaluation was completed within 25 working days by two environmentalists, and while the report is well written it is no stylistic masterpiece. It is not reasonable to expect that it should be; any additional working days would probably be better spent on the substance of report. Even though the report gets a medium rating, it serves its purpose well enough. But if "Sida's Work with Culture and Media" received the same rate, it would not serve its purpose well. The criteria must thus be interpreted with due respect for the diversity among the evaluations.

Annex 4 Terms of Reference

Purpose and Background

During 2005 and 2006, UTV intends to develop models to review the quality of Sida's evaluation system. The work covers quality aspects of evaluation processes – planning and implementation of evaluations – as well as the quality of the finished evaluation reports.

The study also investigates the possibility of compiling and systematically synthesizing the results of Sida's evaluations. Can the results reported in Sida evaluations be aggregated? What are the lessons and operational conclusions of the evaluation system as a whole?

The study is motivated by growing demands for high-quality evaluation information. Over the years, Sida evaluations have been the subject of a number of studies, though many of the studies are no longer topical. Sida lacks a working model for quality reviews that can be used as a basis for regular improvements of the evaluation system

Component Studies

The project has the following components:

1. A quality assessment of Sida's evaluation reports and their terms of reference. This is a desk review of aspects of evaluation quality that are directly accessible through the evaluation report and the accompanying documents. The study focuses on the kind and quality of information presented in the report. What kinds of questions are answered? What kinds of evaluations are conducted? How reliable is the report? The clarity and readability of the reports are also reviewed. Questions regarding the underlying evaluation process and the actual use of the completed evaluation results by stakeholder are not considered.

2. An assessment of the way evaluations are decided, planned, implemented and used in different country contexts. In each country case study the assessment will look at the use of evaluations within the framework of a wider results information system that also includes monitoring mechanisms of various kinds and results information obtained from the national system and other donors. The study will review individual evaluation processes, though it is the system as a whole that is at the centre of the investigation, not the individual evaluation taken by itself.

3. The above-named synthesis study. Here, it is partly about iteratively building a model and partly about testing the model in practice.
4. A summary synthesis with recommendations for developing the activity.

Implementation and Reporting

The study will be conducted by external consultants, in close cooperation with Sida/UTV. Each of the component studies will be preceded by a project description from the consultancy team. The system study will not commence until the results of the other studies have been presented.

The review will be delivered as four separate reports. The scope and design of the reports will be determined during the course of the process. Forms of dialogue with Sida's operational departments, as well as feedback of results to Sida and other stakeholders, will be decided by a reference group set up for the purpose.

Consultancy Support

The project should be conducted by a consultancy team with considerable theoretical and practical experience of evaluation. The project leaders ought to have extensive knowledge of evaluation in international development co-operation and be familiar with research on evaluation quality issues.

In UTV's view, Kim Forss, Andante AB, and Evert Vedung, Uppsala University, together have the right competence for the project. While Kim Forss will be operationally responsible for the study, Evert Vedung's role will be mainly that of an adviser.

Kim Forss's participation is regulated by an existing framework agreement. A new framework agreement will be drawn up with Evert Vedung.

Sida Studies in Evaluation

96/1 *Evaluation and Participation – some lessons.*
Anders Rudqvist, Prudence Woodford-Berger
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=STUD96-1.pdf&a=2368

96/2 *Granskning av resultatanalyserna i Sidas landstrategiarbete.*
Göran Schill
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=STUD96-2.pdf&a=2369

96/3 *Developmental Relief? An Issues Paper and an Annotated Bibliography on Linking Relief and Development.*
Claes Lindahl
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=STUD96-3.pdf&a=2370

96/4 *The Environment and Sida's Evaluations.*
Tom Alberts, Jessica Andersson
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=Stud+96-04.pdf&a=2371

97/1 *Using the Evaluation Tool. A survey of conventional wisdom and common practice at Sida.*
Jerker Carlsson, Kim Forss, Karin Metell, Lisa Segnestam, Tove Strömberg
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=STUD97-1.pdf&a=2366

97/2 *Poverty Reduction and Gender Equality. An Assessment of Sida's Country Reports and Evaluations in 1995–96.*
Eva Tobisson, Stefan de Vylder
Secretariat for Policy and Corporate Development
http://www.sida.se/shared/jsp/download.jsp?f=STUD972.pdf&a=2367

98/1 *The Management of Disaster Relief Evaluations.*
Lessons from a Sida evaluation of the complex emergency in Cambodia.
Claes Lindahl
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=STUD98-1.pdf&a=2363

98/2 *Uppföljande studie av Sidas resultatanalyser.*
Göran Schill
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=STUD98-2.pdf&a=2364

98/3 *Evaluating Gender Equality – Policy and Practice.*
An assessment of Sida's evaluations in 1997–1998.
Lennart Peck
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=STUD98-3.pdf&a=2365

99/1 *Are Evaluations Useful? Cases from Swedish Development Cooperation.*
Jerker Carlsson, Maria Eriksson-Baaz, Ann Marie Fallenius, Eva Lövgren
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=STUD99-1.pdf&a=2355

99/2 *Managing and Conducting Evaluations. Design study for a Sida evaluation manual.*
Lennart Peck, Stefan Engström
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=STUD99-2.pdf&a=2356

99/3 *Understanding Regional Research Networks in Africa.*
Fredrik Söderbaum
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=STUD99-3.pdf&a=2361

99/4 *Managing the NGO Partnership. An assessment of stakeholder responses to an evaluation of development assistance through Swedish NGOs.*
Claes Lindahl, Elin Björkman, Petra Stark, Sundeep Waslekar, Kjell Öström
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=99-4.pdf&a=2394

00/1 *Gender Equality and Women's Empowerment. A DAC review of agency experiences 1993–1998.*
Prudence Woodford-Berger
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=stud00-1.pdf&a=2350

00/2 *Sida Documents in a Poverty Perspective. A review of how poverty is addressed in Sida's country strategy papers, assessment memoranda and evaluations.*
Lennart Peck, Charlotta Widmark
Department for Policy and Socio-Economic Analysis
http://www.sida.se/shared/jsp/download.jsp?f=stud00-2.pdf&a=2351

00/3 *The Evaluability of Democracy and Human Rights Projects. A logframe-related assessment.*
Derek Poate, Roger Riddell
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=stud00-3.pdf&a=2352
http://www.sida.se/shared/jsp/download.jsp?f=stud00-3-2.pdf&a=2352

00/4 *Poverty Reduction, Sustainability and Learning. An evaluability assessment of seven area development projects.*
Anders Rudqvist, Ian Christoplos, Anna Liljelund
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=stud00-4.pdf&a=2353

00/5 *Ownership in Focus? Discussion paper for a Planned Evaluation.*
Stefan Molund
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=stud00-5.pdf&a=2354

01/01 *The Management of Results Information at Sida. Proposals for agency routines and priorities in the information age.*
Göran Schill
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=Stud01-01.pdf&a=2349

01/02 *HIV/AIDS-Related Support through Sida – A Base Study. Preparation for an evaluation of the implementation of the strategy "Investing for Future Generations – Sweden's response to HIV/AIDS".*
Lennart Peck, Karin Dahlström, Mikael Hammarskjöld, Lise Munck
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=Stud01-02.pdf&a=2432

02/01 *Aid, Incentives, and Sustainability.*
An Institutional Analysis of Development Cooperation. Main Report.
Elinor Ostrom, Clark Gibson, Sujai Shivakumar, Krister Andersson
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=Stud02-01.pdf&a=2429

02/01:1 *Aid, Incentives, and Sustainability.*
An Institutional Analysis of Development Cooperation. Summary Report.
Elinor Ostrom, Clark Gibson, Sujai Shivakumar, Krister Andersson
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=Stud02-01Summary.pdf&a=2430

03/01 *Reflection on Experiences of Evaluating Gender Equality.*
Ted Freeman, Britha Mikkelsen
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=44717+UTV+Studies+03-01.pdf&a=2716

03/02 *Environmental Considerations in Sida's Evaluations Revised:*
A follow-up and analysis six years later.
Tom Alberts, Jessica Andersson, with assistance from:
Inger Ärnsfast, Susana Dougnac
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=44818+Stud03-2.pdf&a=2719

03/03 *Donorship, Ownership and Partnership:*
Issues arising from four Sida studies of donor-recipient relations.
Gus Edgren
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=45636+Studies+03-03.pdf&a=2754

03/04 *Institutional Perspectives on the Road and Forestry Sectors in Laos: Institutional Development and Sida Support in the 1990s.*
Pernilla Sjöquist Rafiqui
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=Sida+Eval+03_04.pdf&a=2859

03/05 *Support for Private Sector Development:*
Summary and Synthesis of Three Sida Evaluations
Anders Danielson
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=SIDA3591_UTV03_05_pod.pdf&a=3084

04/01 *Stronger Evaluation Partnerships. The Way to Keep Practice Relevant*
Gus Edgren
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=SIDA4080en_SSE04-01_web.pdf&a=3259

04/02 *Sida's Performance Analyses – Quality and Use*
Jane Backström, Carolina Malmerius, Rolf Sandahl
Department for Policy and Methodology
http://www.sida.se/shared/jsp/download.jsp?f=SIDA4246en_SSE04-02+web.pdf&a=3317

05/01 *Sida och Tsunamin 2004*
En rapport om Sidas krisberedskap
Fredrik Bynander, Lindy M. Newlove, Britta Ramberg
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=SIDA17577sv.pdf&a=12577

05/02 *Sida and the Tsunami of 2004*
– a Study of Organizational Crisis Response
Fredrik Bynander, Lindy M. Newlove, Britta Ramberg
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=SIDA17577en_web.pdf&a=12577

05/03 *Institutionsutveckling skapas inifrån*
Lärdomar från konsulters erfarenheter av stöd till formella och informella regler
Lage Bergström
Sekretariatet för utvärderingar och intern revision
http://www.sida.se/shared/jsp/download.jsp?f=SSE05-03_SIDA23805sv.web.pdf&a=18805

05/04 *Development of Institutions is Created from the Inside*
Lessons Learned from Consultants' Experiences of Supporting Formal and Informal Rules
Lage Bergström
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=SSE05-4_SIDA23805en_web.pdf&a=18805

06/01 *Sida's Management Response System*
Anders Hanberger, Kjell Gisselberg
Department for Evaluation and Internal Audit
http://www.sida.se/shared/jsp/download.jsp?f=SEE06-01_SIDA24695en_web.pdf&a=19695

2007:01 *Erfarenheter av resultatstyrning*
En genomgång av utvärderingar och studier
Lennart Peck
Avdelningen för utvärdering och intern revision
http://www.sida.se/sida/jsp/sida.jsp?d=118&a=31530&searchWords=2007:01

2007:02 *Changing Rules – Developing Institutions*
A Synthesis of Findings
Gun Eriksson Skoog
Department for Evaluation and Internal Audit
http://www.sida.se/sida/jsp/sida.jsp?d=118&a=32203&searchWords=2007:02

2007:03 *'We can't all be ducks'*
Changing Mind-sets and Developing Institutions in Lao PDR
Pernilla Sjöquist Rafiqui
Department for Evaluation and Internal Audit
http://www.sida.se/sida/jsp/sida.jsp?d=118&a=32205&searchWords=2007:03

2007:04 *Evaluations of Country Strategies*
An Overview of Experiences and a Proposal for
Shaping Future Country Programme Evaluations
Stefan Dahlgren
Department for Evaluation and Internal Audit
http://www.sida.se/sida/jsp/sida.jsp?d=118&a=32138&searchWords=2007:04

2007:05 *Mainstreaming at Sida*
A Synthesis Report
Fredrik Uggla
Department for Evaluation and Internal Audit
http://www.sida.se/sida/jsp/sida.jsp?d=118&a=32801&searchWords=2007:05

Are Sida Evaluations Good Enough?

An Assessment of 34 Evaluation Reports

In this study an external team of evaluation specialists takes a searching look at the quality of a sample of evaluation reports commissioned by Sida line departments and Swedish embassies in countries where Sweden is engaged in development co-operation. Assessing the coverage and credibility of the sample reports, the authors seriously question the practical usefulness of the results information generated through Sida evaluations. The report concludes with a set of broad recommendations for improvement.

SWEDISH INTERNATIONAL DEVELOPMENT
COOPERATION AGENCY

Address: SE-105 25 Stockholm, Sweden
Visiting address: Valhallavägen 199
Phone: +46 (0)8-698 50 00
Fax: +46 (0)8-20 88 64
www.sida.se sida@sida.se