Institut für Politikstudien

Seidenhofstr. 12
CH-6003 Luzern
Telefon 041 412 07 12
Fax 041 410 51 82
www.interface-politikstudien.ch

I N T E R F A C E

Quality Assessment of External Evaluation Reports Commissioned by the Swiss Agency for Development and Cooperation

A Case of Evaluation Standards Put to Practice

Paper presented at the 6th conference of the
European Evaluation Society (EES) in Berlin,
September 30 - October 2, 2004

Dr. Luzia Lehmann, lehmann@interface-politikstudien.ch

Dr. Andreas Balthasar, balthasar@interface-politikstudien.ch

Lucerne, 9 September 2004

Table of Contents

Abbreviations and Acronyms

DAC	Development Assistance Committee of the OECD
DAC standards	DAC Minimum Sufficient Evaluation Standards
E & C Division	Evaluation and Controlling Division of the SDC
E & C Net	Evaluation and Controlling Net of the E & C Division
NGO	Non-governmental organisation
OECD	Organisation for Economic Co-operation and Development
SDC	Swiss Agency for Development and Cooperation
SEVAL	Swiss Evaluation Society
SFR	Swiss francs
TORs	Terms of reference
WTO	World Trade Organization

1 Introduction

The evaluation practise of the Swiss Agency for Development and Cooperation (SDC) has undergone considerable change in recent years. The SDC is committed to evaluations constituting a pillar of accountability and learning. In this context, it has established a formal Annual Evaluation Program,[1] which documents that the SDC commissions evaluations regularly today. The SDC's current evaluation program differentiates between the following three evaluation categories:

- Independent evaluations, which are mandated by the Evaluation and Controlling Division (E & C Division) and deal mainly with strategic and policy issues from an outside perspective.
- External evaluations, which are triggered by desk officers supervising operations at headquarters with the aim of assessing approaches and possible alternatives to programs or to thematic areas. These evaluations are conducted within line management and executed by "external consultants".
- External reviews, which are managed by the staff in the head and/or the field offices in charge of program implementation with the aim of improving operations and the orientation of on-going operations.

SDC desk officers commission an average of 30 external evaluations annually, registered in the Annual Evaluation Program. Desk officers lack as yet official evaluation guidelines and standards. The SDC guidelines of 2000[2] or earlier guidelines are not based on the threefold structure of the current SDC evaluation program. At the same time, evaluations in the SDC take place in a challenging environment that requires the cooperation of a chain of people at headquarters as well as in the field. It is in this general context that the E&C Division commissioned a quality assessment with a focus on external evaluations: The SDC wanted to know where it stands with regard to the quality of its external evaluations and at the same time aimed at improving their quality.[3]

The purpose of this quality assessment is thus twofold:

- to render accountability by assessing the quality of external evaluation reports (summative aspect)
- to improve future performance by learning from experience (formative aspect).

The results of the quality assessment should enable the Evaluation and Controlling Net (E & C Net)[4] to better target areas in need of improvement and take measures towards achieving higher evaluation quality throughout SDC. The assessment is not meant to examine the quality of SDC's evaluation program (e.g., relevance of objects under evaluation), the quality of the evaluation function or the dissemination and integration (utilisation) of evaluation results apart from the limited conclusions that could be drawn on the utilisation from the evaluation reports.

The objectives of the evaluation are the following:

- to provide input towards the development of SDC Evaluation Guidelines,

[1] SDC (March 2002-03-04): Ongoing Evaluation Program for 2002-2003 of SDC, Bern
[2] SDC (June 2000): "External Evaluation. Are we doing the right things? Are we doing the things right?", SDC, Bern.
[3] The authors would like to thank Simone Ledermann for her valuable contribution to this study. As part of the assessment team, she analysed six evaluation reports.
[4] The SDC's Evaluation & Controlling Net (E & C Net) is mandated with improving and standardizing evaluations practices as well as with backstopping commissioners of evaluations. The E & C Net comprises the E & C Officers of SDC's six Departments and the staff of the E & C Division.

- to inform the development of SDC evaluation training courses based on the conclusions on the strengths and weaknesses of SDC's external evaluations,
- to contribute to the meta evaluation process of the OECD Development Assistance Committee (DAC) based on the insights gained from the experience made with DAC Minimum Sufficient Evaluation Standards (DAC standards) in terms of their usefulness and through comments on the standards and their applicability in guiding and assessing external evaluations, and
- to contribute to knowledge sharing through better evaluation quality.

The key questions, as outlined in the terms of reference (TORs), were as follows:

1. How does the quality of SDC's external evaluations measure against internationally recognised evaluation standards?

2. What strengths and weaknesses do the evaluation reports, their availability, the TORs, the contracts and the interviews reveal about the commissioning, the execution and the utilisation of the selected external evaluations?

3. How does the quality of SDC's external evaluations measure against the quality of comparable evaluations in other Swiss government agencies?

4. Which level do SDC's external evaluations mainly focus on: the output, the outcome or the impact level? Is the targeted level appropriate measured against the context and against best practices?

5. Do SDC's external evaluations produce information of value that justifies the cost of producing them?

6. What conclusions can be drawn regarding the use of the evaluation results in decision-making?

7. What conclusions can be drawn regarding the appropriateness of the objects of evaluations?

8. What recommendations can be drawn from the quality assessment to improve the use of evaluation results and move in the direction of more focused, more standardised and higher quality external evaluations over the short and the long term?

This quality assessment will tackle the objectives and make recommendations aimed towards improving report quality and, thereby, evaluation impact through the qualitative analysis of a random sample of 12 external evaluation reports and the corresponding TORs and contracts. In chapter 2 the evaluation design and methodology are presented. Chapter 3 presents the results of the assessment for the 12 evaluations and chapter 4 attempts to answer the above key questions by synthesising the results (conclusions). Chapter 5 presents the recommendations and chapter 6 comments on the DAC standards based on their use in this quality assessment.

2 Evaluation Design and Methodology

In this chapter we present the evaluation design including the sample of external evaluations, the list of evaluation criteria established and the methodology.

2.1 Case Study Approach and Sample of External Evaluations

This quality assessment uses a case study approach to answer the key questions: with a qualitative analysis of a random sample of twelve external evaluations conducted in 2002.[5] Although a case study approach does not lend itself to quantitatively based generalisations, such an approach will reveal the major patterns both in terms of strengths and weaknesses prevalent in external evaluations.

2.2 Standards of Evaluation

The quality of evaluations has to be assessed by applying a set of standards. The starting point for devising such a set were the DAC Minimum Sufficient Evaluation Standards. The SDC as the commissioning body wanted the DAC standards to be used as the basis for this assessment in order to obtain a feedback on the utility and applicability of the DAC standards. As a member of OECD Development Assistance Committee (DAC), Switzerland wanted to contribute to the meta evaluation process of the DAC.
The DAC standards - quality standards for evaluation reports in development assistance – are based on a review of the "Principles for Evaluation of Development Assistance"[6] and other internationally recognised standards such as those of the Swiss Evaluation Society and the American Evaluation Society.

For the purposes of this quality assessment, we complemented the DAC standards with the widely used SEVAL standards established by the Swiss Evaluation Society. The list of standards for comparison with actual values was thus established on the basis of:

- the DAC Minimum Sufficient Evaluation Standards,[7]
- the SEVAL standards established by the Swiss Evaluation Society,[8] and
- key questions of the Terms of Reference if these were not covered by the above.

We thus devised a list with 23 standards divided into four categories.[9] We rely on the four widely used categories of standards from SEVAL rather than the six from the DAC standards. The problem of overlap, which is inherent in standards and categories anyway,[10] can thus be minimised. Moreover, we find the SEVAL categories more focused and to the point, more manageable, easier to delineate and thus easier to apply for the purposes of a meta evaluation. The categories can be summarised in straightforward words as follows. The category *utility* refers to readable, accessible and timely evaluations with a good and useful summary. *Feasibility* ensures that an evaluation is executed in a realistic, well-thought out manner. *Propriety* deals with the ethical aspect and requires, for example, that what is left out also be made explicit. *Accuracy* ascertains that proper scientific methods and procedures are used.

The standards were then filled in a table and each one of them was accompanied by a set of questions fleshing out the meaning in an effort to apply the standards. For each evaluation the list of criteria was worked through systematically and the findings filled in a fact sheet. The resulting 12 fact sheets of the case studies were then compared synthetically (chapter

[5] SDC (March 2002-03-04): Ongoing Evaluation Programme for 2002-2003 of SDC, Bern. Evaluations were selected from each SDC department, with the exception of Department Services (which did not conduct any in 2002) and with a stronger weighting of the 3 operations departments conducting more evaluations (Department for Bilateral Cooperation, Department for Cooperation with Eastern Europe, and Department for Humanitarian Aid). There is one evaluation each from the Department for Development Policy and Multilateral Cooperation and the Thematic Department. The sample includes 2 thematic evaluations, 5 institutional evaluations and 18 project evaluations.
[6] Development Assistance Committee (1991): "Principles for Evaluation of Development Assistance", OECD, Paris, 1991.
[7] See appendix.
[8] http://www.seval.ch/en/standards/index.cfm.
[9] See appendix.
[10] Widmer, Thomas (1996): Meta-Evaluation, Kriterien zur Bewertung von Evaluationen, Bern, Haupt.

3). Finally, on the basis of the results of this comparison, conclusions (chapter 4) and recommendations (chapter 5) were formulated.

2.3 Methodology

In order to assess the evaluations and fill in the fact sheets, the following sources were used:

- We reviewed the relevant and available documents: evaluation reports, terms of reference (TORs), contracts, budgets and final statements, and SDC evaluation guidelines and instruction manuals, if applicable.

- We conducted personal interviews with the desk managers of the external evaluations, in several cases both former and later/current desk managers (14 interviews).

- Where necessary and possible we conducted interviews with the evaluators (10 interviews).

3 Assessment of External Evaluations

In this chapter we present the findings of the quality assessment of the twelve evaluations. The evaluations were assessed against a set of 23 standards grouped into the four categories utility, feasibility, propriety and accuracy. Each standard was fleshed out with questions that helped apply the standards. This resulted in comments and assessments for each standard filled in separate fact sheets for each evaluation. For each standard we also attempted a shorthand assessment on a scale of five levels.[11] Working through the twelve evaluations with all the standards in this manner thus resulted in twelve fact sheets.

The structure of this chapter follows the order of the categories and standards. For each standard we first list the questions that helped apply the standards and then comment on all the standards applied to the sample of evaluations, illustrating strengths and weaknesses.

A caveat on the assessments in general is in order: The twelve evaluations in the sample constituting the reference group for this analysis are all short-term and fairly small or small evaluation projects.

3.1 Utility

The utility of evaluations refers to readable, accessible and timely evaluations with a good and useful summary. This category was measured with the following 11 standards.

3.1.1 Stakeholders Identified (1)

Does the report identify the ultimate beneficiaries of the evaluation, the core learning partners and those best positioned to implement the recommendations?

The evaluations get a fairly good rating on this standard. Half the evaluations fulfill the standard, 5 mostly and 1 in part. Stakeholders were generally identified fairly well in the evaluations. Nevertheless, they were rarely presented in a systematic manner listing the categories at the outset, with a view to easy readibility and accessibility. Sometimes the

[11] The following scale was used: +: positive ("fully fulfilled"); (+); rather positive ("mostly fulfilled"); +/-: neutral ("in part fulfilled"); (-): rather negative ("mostly not fulfilled"); -: negative ("not fulfilled/considered"); n.a.: no assessment possible due to missing or incomplete information. In order to ensure a degree of comparability, we use the same categories as Widmer (1996, p. 242).

various stakeholders appeared in the course of the report in a piecemeal fashion and at times even under various names. In some cases, there was a complete list in the appendix.

3.1.2 Evaluation Purpose and Objectives Clear (2)

Do the TORs and the report clearly state the primary purpose and the objectives of the evaluation? Is the process adopted to ensure that all stakeholders understand the objectives described? Is the level of evaluation (output, outcome, impact etc.) clearly stated and appropriate?

In the majority of cases the degree of fulfilment of this standard is on the problematic side. Only three evaluations mostly fulfil the standard, whereas 4 do so in part, 4 mostly do not fulfil it and 1 does not fulfil it. There were the following shortcomings on the issue of *clear purpose and objectives*. The overall purpose was not always clear and/or not linked to the objectives and key questions coherently enough. It also happened that that the purpose outlined in the TORs and that in the report were not identical. In some cases there were too many objectives or the level of detail was excessive in that there were key questions with a large number of sub-questions. One evaluation with a low rating had no clear overall purpose at all and the key questions given were not linked to the sub-questions.

The consequences for the evaluators when this standard was poorly fulfilled were manifold:

- First, confusion about the TORs among the commissioning division within SDC: it became clear during the process that the objectives were not clear. In one case, since the SDC divisions involved in drafting the TORs and commissioning the evaluation failed to agree, the evaluator was unable to fulfill competing and unclear objectives, was left dangling and the evaluation project ended up as a failure, terminated prior to the second mission.
- Second, the TORs were not understood properly by the evaluators. In one case the result was two separate evaluations pursuing different objectives and producing separate evaluation reports. In a second case the evaluator successfully suggested a discussion to clarify the unclear objectives and reduce their large number. Finally, in another case the same problem was not resolved but postponed, therefore resulting in a report with comments by the evaluator such as “what did you actually mean?”

Almost all evaluations fail to explicitly *describe the process* adopted to ensure that all stakeholders understand the objectives. However, in the interviews it became clear that such a process often takes place orally. Since the evaluations are typically participatory in nature and the involvement of stakeholders is common practice (standard 11), this does not seem to be a problem.

All the evaluations had a focus on the *output level of evaluation*, which attempts to assess services provided by a project or program. The *outcome level* assessing a project or program’s influence on the immediate target groups was also focused on by all evaluations. The *impact level*, assessing the socio-economic and socio-political effects, was a focus of three quarters of the evaluations at least to some degree. The levels the sample of external evaluations are to focus on are at times explicitly and clearly mentioned. At times they become (more or less) clear only from the description of tasks, so they are only implicitly stated. In some cases it appears that the terms are used rather like buzz words, as questions at the impact level are required in TORs, yet are in most cases impossible to evaluate systematically in such small projects. Assessing the impact level makes sense only for projects and programs that have been running for a while and for evaluations of a certain size. Therefore, in evaluations of longer-running projects, the focus on outcome and impact levels would be appropriate for an evaluation in theory, yet hardly for small evaluations with low financial and small time budgets, as most of those in the sample assessed here.

3.1.3 Demand Responsive (3)

Is the evaluation focused on the central questions of the TORs and does it answer them all and in a way that reflects their stated level of priority? Are the recommendations useful? Is the object of evaluation appropriate?

Overall, the assessment of the evaluations on this standard is slightly positive: 9 evaluations got a rating of either positive (3) or rather positive (6), 2 neutral and 1 rather negative. In several instances where not all *key questions* were answered in full or in-depth, this was explicitly said - in the report and/or in the interviews - to have been due to unrealistic TORs. Moverover, several reports follow the questions in the TORs slavishly, which does not always make for easy reading. On the whole, there is need for improvement in the following areas.

Recommendations are considered useful in the majority of evaluations. However, they are not always clearly separated from results or conclusions. Some recommendations are mixed up with the presentation of the results and can therefore not be found easily because of their location. The evaluation with the lowest rating on this standard lacks explicit recommendations in the report, but conclusions are in fact recommendations. Some evaluations suffer from a large number of recommendations that are neither grouped nor prioritised. In one case the wording of the recommendations was too diplomatic, as the desk manager put it.

The o*bject of evaluation* is generally appropriate in theory but in several cases on the broad side given the small size of the evaluation projects (e.g. one small evaluation was supposed to evaluate seven projects), hence it was mostly the breadth of the object of evaluation that resulted in incompletely or not systematically answered questions.

3.1.4 Demonstrated Professionalism and Competence (4)

Can the evaluation team be considered as credible regarding its qualifications and experience in the evaluandum?

From our limited perspective it is difficult to judge whether an evaluator had the necessary qualifications and experience. However, we can say that there are no indications for a lack of qualifications and experience. Desk managers gave the evaluators fairly high marks for their professionalism and competence: 8 get top marks, 3 get near-top marks. Only one evaluator's competence was severely questioned by the desk manager. Yet part of the harsh judgment may be due to the highly difficult circumstances of the failed evaluation process because in the selection procedures the evaluator had been deemed competent.

3.1.5 Selection Procedure of Evaluation Team (5)

Was the evaluation team appointed directly or was there a competition? Were those responsible for the evaluandum able to influence the choice?

The assessment of the *selection procedures* of the evaluation teams turned out to be mostly critical: 2 evaluations fulfilled this standard in part, whereas 5 were rated either rather negative or negative. The reason is that all twelve evaluators or evaluation teams were *appointed*, there were no full-fledged competitions. What was also taken into account in the assessment was the size of the project: The higher the budget of the evaluation, the more negative the assessment if there was little or no competition. The other aspect – influence on the selection procedures by those responsible for the evaluandum - was not considered in the assessment (see below). Nevertheless, it needs to be kept in mind that a negative assessment of the selection procedure says nothing about the competence of the evaluator.

Often the exact selection process cannot be recounted so that the procedures appear insufficiently transparent as to how many candidates were discussed, which stakeholders suggested which candidates, why exactly he or she ended up being chosen etc. In hindsight, therefore, it has to be said that the procedures are not transparent. The most common procedure was the following: The desk manager decides what his or her needs for the evaluation are and thinks of the necessary criteria, such as thematic and regional know-how, experience in the geographical area or with certain institutions, international or national reputation in the field, confidence among stakeholders, short-term availability etc. Then the desk manager may make a shortlist, seeking suggestions from other SDC divisions, the SDC coordinating office abroad, implementing agencies or other stakeholders. In several cases, several candidates were considered and discussed – although the exact number of those considered and the decision-making process can often not be reconstructed in full. It seems that in about half the cases only the evaluators eventually appointed were considered in the process.

Those *responsible for the evaluandum* were consulted in several instances or had a say with regard to the choice of evaluators. The desk managers concerned consider this a success factor for evaluations. In some cases it took some convincing on the part of SDC because the implementing agency of the program under scrutiny initially rejected the choice. From a learning perspective, the involvement of those responsible for the evaluandum in the selection of the evaluator is a good thing. Such involvement is indeed general practice in Swiss federal offices; this is not the case with supervisory and auditing bodies such as the Swiss Federal Audit Office (Eidgenoessische Finanzkontrolle) and the Parliamentary Control of the Administration (Parlamentarische Verwaltungskontrolle). However, it is a question of the degree of influence as to whether the issue of independence (problem in terms of accountability) is at stake. A balance between a learning perspective and accountability needs to be found. In one case the implementing agency (1 person) suggested the local evaluator who was then appointed to the evaluation team: this person, however, seemed to have been insufficiently independent and too close to the evaluandum.

An additional finding with regard to the selection of an evaluation team was the following: *Cooperation and division of labor among the evaluators on an evaluation team do not ensue as a matter or course*. Some teams functioned beautifully, even when evaluators did not know each other beforehand, others did not. One team fell apart before it materialised. The tight time schedule typical of SDC external evaluations aggravates the issue. A careful selection and matching of evaluators is necessary, as is a thorough briefing both in terms of evaluation content, coordination, division of labor and responsibility in the team. In some cases the TORs clearly define the coordination and division of labor.

3.1.6 Comprehensive and Clear Reporting (6)

Does the evaluation report precisely describe the object of evaluation? Is the evaluation report logically structured and does it outline the evaluation context, goals, questions posed, and procedures used, as well as any constraints encountered that substantively hindered its ability to fulfill its purpose? Is there an executive summary with key findings, conclusions and recommendations?

The record on comprehensive and clear reporting is mixed, with 3 reports getting the top rating, 4 a rating of rather positive, 2 neutral, 2 rather negative and 1 a negative rating. The assessment takes into account the small or fairly small size of the 12 evaluations and is thus on the “benevolent” side. The following comments focus on the room for improvement of the quality of reports. On the whole, evaluation reports tend to be written for a narrow audience and are thus not easily accessible to someone unfamiliar with the evaluandum. Stripped of appendices and summaries, the reports are between 10 and 46 pages long.

In several reports the *object of evaluation* and the *evaluation context* are described cursorily, insufficiently clearly or in a piecemeal fashion. At times they were simply copied from the TORs in whole or in part. In one case there was no description of the object altogether. The *goals* of the evaluation were generally better presented than the *questions posed.* The latter were in some cases not explicit enough and/or referred the reader to the annex. The description of the *procedures used* was in several instances too general, with references to the itinerary, interviews or field visits made or the use of a participatory approach (see standard 18). In those reports where *constraints encountered* were described, the information provided was helpful. Such information included issues like the following: time limits of the missions or unexpected complexity of an object making it impossible to fully answer certain (aspects of) questions, people unavailable for interviews and the consequences for the analysis, the political situation making field visits impossible and how the problem was dealt with, difficulties with insufficient data as well as insufficiently comprehensible data.

The quality of the *structures* of the reports varies considerably. The good ones are logical in structure, make the structure explicit at the outset, clearly separate the various projects or program aspects considered as well as findings, conclusions and recommendations and include transitions between the various parts that make it easy for the reader to follow the analysis and argument. Some reports, however, fail to separate the parts well enough for the reader to follow. This is the case when parts of the report fail to cover all the projects under consideration in an evaluation. In one case only one of four projects evaluated is mentioned in the conclusion. Others included repetition or even contradiction on certain subjects. In some cases the repetition or unsatisfactory structure is due to the fact that the TORs are followed slavishly.

Readability is hampered in those reports that exhibit *shortcomings in formal layout*: no page numbering, wrong chapter numbering resulting in confusion as to what project is under discussion. Some reports lack a table of content, or have one that is inconsistent with the text of the report, with chapters missing or without page numbering. More than half the reports are fairly strong in language usage, i.e. spelling and grammar mistakes are minor and do not impede understanding. Yet three reports are in need of considerable editing. Problems of unclear references (“this”, “it”) and tenses (unclear whether empirical findings or recommendations are presented) provide the potential for misunderstanding. Needless to say a text with a lot of formal shortcomings frustrates the reader and possibly leaves him or her questioning the credibility of the report in general.

Executive summaries (between 1 and 10 pages long) are available for 11 evaluations, with considerable variations in quality. While some are highly useful and include key results, conclusions and recommendations in a well structured manner, others leave out parts or, worse yet, are not congruous with the content of the report, which happens when recommendations are grouped or prioritised differently and, as in one case, are not identical. Only 4 DAC abstracts are available, even though SDC standard operating procedure requires them for external evaluations.

3.1.7 Transparency of Value Judgments (7)

Are the underlying reasoning and points of view upon which an interpretation of evaluation results rests described in such a manner that the bases for the value judgments are clear?

On the whole, the transparency of value judgments received a slightly positive assessment: 3 reports get a positive and 6 a rather positive rating, 1 gets a neutral and 2 a rather negative rating. The shortcomings include: insufficient separation of the presentation of findings, analysis, conclusions and recommendations. Several reports remain largely descriptive,

amply presenting results and quotations from interviews but little analysis and interpretation. One report presents conclusions, yet virtually no findings and no analysis. In two reports it is occasionally unclear whether a general statement is being made or empirical results are being presented, or whether empirical results or the recommended state are presented. In interviews with desk managers it became clear that those desk managers who did not accompany an evaluation from the start but "inherited" it later were more critical on this point than the others.

3.1.8 Timely Reporting (8)

Have significant interim results as well as the final report been made available to the intended users in such a way that they can be utilised in a timely manner?

The evaluations considered are generally fairly good on making the final report available to the intended users in a timely manner. Most evaluations were characterised by tight time schedules. Some evaluations were done in such short time spans that the availability of interim results was hardly possible. Interim results consisted mainly of drafts of (final) reports on which SDC and in several cases other stakeholders provided feedback.

In three cases with delays in the final reports, the delay did not impede the use of the evaluation according to the desk managers. In one case the delay was due to difficulties in scheduling the debriefing; the contract was amended to reflect the delay accordingly. In the same case, some stakeholders (e.g. the implementing agency) were critical because they did not receive the draft report prior to the debriefing, whereas the desk manager argued it was general practise not to disseminate preliminary results to the outside. In one case the evaluators did not get any feedback on the draft report so that their draft became the final report. The evaluation with the lowest rating on timely reporting was unable to progress as planned and was terminated before the second mission started: The two SDC divisions involved did not agree on the (unclear) objectives in the TORs so that the evaluator could not do justice to both divisions and had to write several drafts, yet to no avail.

3.1.9 Evaluation Impact (9)

Do the planning, execution and presentation of the evaluation encourage stakeholders to follow the evaluation process and to utilise the evaluation? Are the various interests taken into account in order to win their cooperation?

Overall, this standard is assessed as fairly satisfactory, yet there is some room for improvement: 4 evaluations are assessed positive and 5 rather positive on evaluation impact, whereas 2 are rated in part fulfilled and 1 rather negative.

As far as *planning* is concerned, the coordination and mutual understanding of the purpose and objectives of the evaluation among those commissioning and using the evaluation is crucial. The early information and consultation if not involvement of relevant stakeholders in order to obtain confidence in the evaluation process, especially when sensitive issues are at stake, is also relevant. In the majority of cases this stage was fairly well managed. Where this was not the case, the conditions for the utilisation of the evaluations were negatively affected, as the following examples show. In one case with problems at the planning stage, the two SDC divisions commissioning the evaluations (and NGOs) did not see eye to eye on the main objectives and were unable to draft precise and unambiguous TORs. As a consequence, the evaluator could not fulfill expectations of both divisions, unable to revise drafts satisfying both divisions. The evaluation was then terminated after only the first of two planned missions. In two other instances, there seemed to have been - unspoken - different political understandings of whether the change of or end of funding for an implementing

agency was admissible. In one case the result was that the desk manager in charge of implementing the recommendations of an evaluation done previously was not supported by senior management in SDC in implementing the recommendation to discontinue funding for an implementing agency.

The *execution* was rather positive for a majority of evaluations insofar as stakeholders participated in the process (see standard 11), desk managers and/or SDC coordinating offices accompanied the evaluations and evaluation teams cooperated well. On the one hand evaluators thus got support and feedback. On the other hand a learning process for (SDC) stakeholders took place. This was particularly apparent in one case, where the representative of the coordinating office accompanied the evaluation very actively in an area where the SDC had little know how and competence and thus improved the latter considerably. The problems at the execution stage present in some evaluations involved, first, insufficient continuity on the part of SDC in accompanying an evaluation. This was mainly due to the fluctuation of personnel among desk managers and coordinating offices and occasionally left evaluators without feedback during the process. For example, in one instance the evaluator should have assessed an aspect of the evaluandum more critically, as the former and later desk managers agreed, but the first desk manager left right after the TORs were drafted and the successor did not arrive until after the evaluation was finished and the recommendations were supposed to be implemented. A second obstacle in the execution of several evaluations concerned insufficient or a lack of cooperation among the members of the evaluation teams. The main reasons for problems at this level were disagreements about the objectives of the evaluation and the unclear division of labor among the evaluators.

The *presentation* stage generally involved a debriefing with the major stakeholders, either at SDC headquarters or for smaller evaluations in some instances only at the SDC coordinating office abroad. In two cases the SDC failed to provide feedback to the evaluators at all. The evaluation with the lowest rating on providing conditions conducive to utilisation did poorly on all aspects above: stakeholders received little information at short notice about the planned evaluation and were not involved in the execution; the whole process including the presentation was largely an SDC-internal affair.

On the whole, the conditions for the *utilisation* of evaluations were fairly good for more than half the evaluations. In the vast majority of cases the various *interests* were taken into account in order to win their cooperation.

3.2 Feasibility

The second group of standards deals with the feasibility of evaluations. Feasibility ensures that an evaluation is executed in a realistic, well-thought out manner. This category was measured with 3 standards.

3.2.1 Practical Procedures (10)

Are evaluation procedures designed in a way that the information needed is collected without unduly disrupting the object of the evaluation?

This standard received the strongest assessment of all and thus seems to be a strength of SDC external evaluations. Since external evaluations typically use a participatory approach relying on interviews with the stakeholders, it is not surprising that such procedures are not considered disrupting and that evaluators have the necessary experience and social competence for such an approach.

3.2.2 Evidence of Participation (11)

Did the stakeholders have the chance to participate and introduce their views?

In 9 evaluations the evidence of participation is positive. In two evaluations this standard is mostly fulfilled: In one of these a group of stakeholders (abroad) was not asked to participate because project funding was too low to include them (the TORs are contradictory on this point: They speak of considering "all partners", yet the project does not provide the resources to do so). Another evaluation was to a large degree an SDC-internal affair. The evaluation that fulfilled the standard only in part is the one exhibiting disagreement among the SDC divisions involved: the evaluator one-sidedly leaned to more participation on the part of the SDC division that was more precise on interests and objectives.

3.2.3 Costs and Cost Effectiveness (12)

What were the costs of the evaluation? Does the evaluation produce information of a value that justifies the cost of producing them (value for the money)?

According to the information available, the *cost* of the 12 evaluations was in the range of SFR 7'500 to SFR 63'000. However, the total and exact cost in Swiss Francs is transparently documented only for 7 evaluations. For the remaining 5 the documentation available either includes budgets and statements for only part of the evaluators on a team (often the local evaluator is missing, in one case at least indicated in local currency). In one case the budget had to obtained from the evaluator himself, in another it was missing entirely. In addition the final statement is missing for 5 evaluations (in two cases, oral information was provided by SDC).

For the assessment of the *cost effectiveness* we relied largely on the desk managers' views. In general the cost effectiveness of the evaluations was judged positively, and SDC seems to have received value for the money, 6 evaluations getting a positive and 5 a rather positive assessment. The evaluation with the rather negative assessment was the one terminated prematurely. Nevertheless, a caveat is in order regarding the objectivity of the assessment of the cost effectiveness. The evaluations under consideration tend to have been small projects with low budgets and are thus by nature more easily seen as cost effective. Indeed one desk manager said the cost effectiveness was more due to the very low cost in absolute terms than on the actual utility of the evaluation. Moreover, the desk managers – the main source for this assessment – provide the perspective of a stakeholder, user and commissioner and thus a reflection of how useful the evaluations were to them. Their assessment fails to provide an independent political perspective. However, the low level of transparency as to the cost is problematic at any rate.

3.3 Propriety

The propriety of evaluations deals with the ethical aspect of evaluations and requires, for example, that what is left out also be made explicit. This category was measured with 4 standards.

3.3.1 Formal Written Agreement (13)

Is there a formal written agreement specifying the duties of the parties who agree to conduct an evaluation? Does this agreement clearly state the areas to be addressed by the evaluation [scope of work], the key questions, the resources and the time allocated, methodology and procedures to be followed, and reporting requirements?

The evaluations reveal some deficiencies regarding the standard formal written agreement, as 2 evaluations are assessed “positive”, 5 “rather positive” and 5 “in part fulfilled”.

The *formal written agreements* (contracts) specifying the duties of the parties are generally clear, yet their availability is mediocre. Eventually at least one agreement per evaluation team was provided, but several agreements of additional evaluators on the teams remained missing. For one evaluation the agreement had to be obtained from the evaluator as SDC did not have one. In one case all three agreements with the evaluators lacked dates, signatures. Several contracts were signed weeks after work on the evaluation had begun. (As for the *resources* allocated – budgets and financial statements - see standard 12 above.) The *division of labor and responsibility* for the evaluation team is well described in several TORs and insufficiently in others. In one case there was mention of who was in charge, yet the point was not communicated successfully by the commissioner: The team never materialised and ended up conducting two separate evaluations and reports (one of which was included in the sample here).

As far as the TORs are concerned (see also standard 2), few nicely outline the *areas to be addressed, key questions, and time* allocated as well as the steps concerning *methodology and procedures*. In many, the information is not clear enough, incomplete or parts missing. As for the description of *methodology and procedures*, some TORs had no details whatsoever, others included a list of information sources to consult, possible interview partners or a reference to a participatory approach. *Reporting requirements* range from clear and comprehensive to vague (e.g. maximum pages).

3.3.2 Complete and Balanced Assessment (14)

Is the evaluation complete and balanced in presenting and assessing the strengths and weaknesses of the object being evaluated?

In 7 evaluation reports including comments by interviewees we conclude that they are complete and balanced in their assessment as there is no indication of the opposite. In 1 case, certain aspects seem to have been left out, in another there were some contradictions and a third one seemed insufficiently critical of the evaluandum (the desk managers agreed with the authors). One evaluation was judged as being excessively critical on an aspect and incomplete in another, though the desk manager attributed part of the problem to the TORs. The evaluation that was terminated prematurely could not be assessed on this count.

3.3.3 Making Findings Available (15)

Are the results made available to all the potentially affected persons as well as to all others who have a legitimate claim to receiving them?

No evaluation fulfilled this standard completely, 9 mostly fufilled it, 2 in part and 1 mostly did not fulfill it. This standard can be narrowly or broadly conceived as far as the circle of those who have a legitimate claim to receiving the evaluation reports is concerned (see chapter 5). This assessment is “middle of the road”. The effort and time to obtain materials as well as certainty on their status (Is it the final report? Is the summary, the DAC abstract, the appendix complete?) was extraordinarily high even though the authors doing the quality assessment had the support of SDC staff.

3.3.4 Declaring Conflicts of Interest (16)

Are conflicts of interest addressed openly and honestly so that they compromise the evaluation process and conclusions as little as possible?

In several cases there was indeed evidence of such conflicts, and they were addressed explicitly in the reports and the problem thus appropriately handled. For one evaluation this standard could not be assessed: No conflicts are addressed in the report, yet the quality of the report is poor and the desk manager does not think much of the quality of the evaluation either. In four cases there was indication that some potentially conflictive issues should have been addressed. The report with the lowest rating on this standard did not address conflicts of interest, even though severe conflicts appeared between the commissioners while the evaluator was unable to fulfill the mandate given the problematic TORs. For 6 evaluations we have no indication that conflicts of interest were not openly and honestly described.

3.4 Accuracy

The accuracy of evaluations ascertains that proper scientific methods and procedures are used. This category was measured with 7 standards.

3.4.1 Identifying and Analysing the Context (17)

Are the influences of the context on the object of evaluation identified and described?

Two thirds of the evaluations fulfilled this standard, identifying the influences of the context on the object of evaluation. In two reports the context was addressed too cursorily, in another one cursorily as well as in a piecemeal and totally unsystematic fashion. Nor did the latter three make explicit what influence the context had on the object of evaluation. One evaluation did too little to fulfill this standard.

3.4.2 Precise Description of Evaluation Procedures (18)

Is there a detailed description of the organisation of the evaluation, data collection and processing, analysis and reporting? Are the procedures used sufficiently precisely described and documented so that they can be identified as well as assessed? Is the choice of method discussed in the report?

We assessed the presence of clear information on methods, procedures, analyses and thus whether the bases for assessments in the evaluations were present ("Nachvollziehbarkeit"). This standard is not a strength in the sample of external evaluations. Only two reports fulfil this standard in full, whereas 3 mostly fulfill it, 5 fulfill it in part, one mostly not and one not at all. Several desk managers were also critical of the reports in this respect, though at times only mildly, yet maintaining that they do not want reports showing that evaluators are "evaluation method cracks".

Weaknesses in elaborating *the organisation of the evaluation, data collection and processing, analysis and reporting* mean that several of the following elements were missing or incomplete: step-by-step descriptions of the evaluation process, mission itineraries and lists of interviewees, description of data collection instruments such as structured interviews or written surveys with questionnaires, some information on data analysis, definition of indicators used for assessments (or even on how many projects analysed the presented results were based). One evaluation included quantitative analyses that presented the results in a manner difficult to understand, with explanations spread out throughout the report. Another weakness of this evaluation was the inclusion of material on methods in the annex yet insufficient information on their significance or use. Most reports were without mention of scales used in assessments. Some reports included little analysis or assessment by the evaluator, presenting results for example in the form of quotations from interviews without discussing their relevance. Other reports used vague concepts such as the reference to a

"systemic approach to research". The evaluation with the second lowest rating included no documentation of procedures used and the one with the lowest rating no documentation at all except for a list of people who accompanied the evaluator.

The *choice of method* and its limits in the context given was discussed only in one evaluation.

3.4.3 Trustworthy Sources of Information (19)

Are the sources of information used in the evaluation sufficiently precisely described so that their adequacy can be assessed?

5 reports fulfilled this standard, 2 mostly, 3 in part, 1 mostly not (with no mention of documents or interview partners) and 1 not at all (indicating only one source). The five reports in the middle range did not list all interview partners including institutional affiliation and a list of documents consulted including references in the text regarding the sources used. In a couple of cases such information might have been in the appendix, yet if the appendix was not available to us, the information was assessed as missing.

3.4.4 Valid and Reliable Information (20)

Are the data collection instruments selected, developed and employed valid and reliable? Are methods and procedures applied as stated and in accordance with their own quality standards?

Validity is determined by assessing the degree to which the instruments employed accurately reflect the concepts they are intended to measure. *Reliability* refers to the consistency or stability of the quality measured, whether between measurement instruments, persons, or over time. We were unable to gauge the standard properly with the information at hand. This is not surprising given that the evaluation reports tend not to be generous with information of procedures.

3.4.5 Impartial and Substantiated Conclusions (21)

Are the conclusions reached in the evaluation clearly and explicitly described and substantiated in such a manner that stakeholders can comprehend and judge them?

5 evaluation reports fulfilled the highest rating on impartial and substantiated conclusions (see also standard 14), 3 mostly fulfilled this criterion, 2 in part, whereas one mostly did not. Concerning the evaluation that was terminated prematurely, it was impossible to assess this standard objectively on our part, whereas the desk manager interviewed both totally disagreed with the evaluation's conclusions and challenged their bases. One evaluation that was rated in part fulfilled came across as excessively positive in some of its assessment. This may have been fully appropriate, but the author failed to substantiate the statements (a view shared by the desk managers). The author did not indicate how many of the projects surveyed produced a certain finding, and some of his assessments were too far-reaching for the thin empirical evidence provided. Another report left the reader in part unsure whether some statements were based on empirical evidence or whether they constituted a recommendation.

3.4.6 Neutral Reporting (22)

Is the report free from distortion through personal feelings or preferences on the part of any party to the evaluation? Does the evaluation report present conclusions in a neutral manner?

Eight evaluation reports are fully neutral in reporting, 2 mostly neutral and 2 in part neutral. One of the latter evaluations included some condescending wording, another some reproachful statements and very emphatic (though generally not offensive) language. A third one had a reproachful undertone at some points ("you could have saved time, anger and money"). A report that was mostly neutral had some excessively metaphorical and exaggerated language that seemed out of place.

3.4.7 Enabling Metaevaluation (23)

Is the empirical material of the evaluation available in order to enable a meta evaluation to check if the evaluation is appropriately executed so that stakeholders can assess the evaluation's strengths and weaknesses?

The evaluations assessed do not lend themselves easily to meta evaluation. Only 2 evaluations fulfill this standard in full, 4 mostly fulfill it, 5 in part and 1 mostly does not. Aspects that need to be included in the assessment of whether a meta evaluation is possible are the quality as well as the availability of the data basis. We have elaborated above on the difficulties associated with obtaining relevant documents.

4 Conclusions: Strengths and Weaknesses of SDC's External Evaluations

In this chapter we present the conclusions based on the assessment of the twelve external evaluations discussed in the last chapter.

4.1 The Quality of External Evaluations Reports

The table below captures an overall assessment of the quality of the evaluation reports, yet a caveat is necessary. Assigning a shorthand assessment to each standard is a difficult endeavor, one of a qualitative rather than a scientific nature. The results depict a general trend of our analysis; they do not provide the basis for a quantitative analysis.

The evaluations in the sample get the highest marks in the *category feasibility*. The evaluations thus exhibit a strength in designing procedures that do not unduly disrupt the object of evaluation (practical procedures), in ensuring the participation of stakeholders and in the evaluations' cost effectiveness.

The record on the other three categories is more ambiguous, including strengths and weaknesses side by side. The assessments of the standards in the *category propriety*, which deals with the ethical aspect of evaluations, show the following mixed picture: Evaluations tend to be on the stronger side in making complete and balanced assessments, yet when it comes to formal written agreements, declaring conflicts of interest and making findings available, they tend to exhibit weaknesses rather than strengths.

The range from strengths to weaknesses is even more pronounced in the *category utility*. Here, the evaluations have their strengths in identifying stakeholders, the professionalism and competence of evaluators (an assessment largely based on desk managers' views) and timely reporting. The weaknesses of the evaluations in this category include the standards clarity of evaluation purpose and objectives, the selection procedures of the evaluation team as well as comprehensive and clear reporting. The reports tend to be written for too narrow a target group of desk managers, coordinating offices, implementing agencies and occasionally other stakeholders directly affected. Yet the reports may be of interest to a wider

audience such as SDC management, political actors, foreign development agencies or NGOs.

Table 4.1: Overview on Assessments of the 23 Evaluation Standards

Category	Standard	+	(+)	+/–	(–)	-	n.a.
Utility	*1 Stakeholders identified*	6	5	1	-	-	-
	2 Evaluation purpose and objectives clear	-	3	4	4	1	-
	3 Demand responsive	3	6	2	1	-	-
	4 Demonstrated professionalism and competence	8	3	-	-	-	1
	5 Selection procedure of evaluation team	-	-	2	5	5	-
	6 Comprehensive and clear reporting	3	4	2	2	1	-
	7 Transparency of value judgements	3	6	1	2	-	-
	8 Timely reporting	8	2	1	1	-	-
	9 Evaluation impact	4	5	2	1	-	-
Feasibility	*10 Practical procedures*	12	-	-	-	-	-
	11 Evidence of participation	9	2	1	-	-	-
	12 Costs and cost effectiveness	6	5	-	1	-	-
Propriety	*13 Formal written agreement*	2	5	5	-	-	-
	14 Complete and balanced assessment	7	3	1	-	-	1
	15 Making findings available	-	9	2	1	-	-
	16 Declaring conflicts of interest	6	1	3	-	1	1
Accuracy	*17 Identifying and analysing the context*	8	2	1	1	-	-
	18 Precise description of evaluation procedures	2	3	5	1	1	-
	19 Trustworthy sources of information	5	2	3	1	1	-
	20 Valid and reliable information	-	-	-	-	-	12
	21 Impartial and substantiated conclusions	5	3	2	1	-	1
	22 Neutral reporting	8	2	2	-	-	-
	23 Enabling metaevaluation	2	4	5	1	-	-

+: positive ("fully fulfilled"); (+); rather positive ("mostly fulfilled"); +/-: neutral ("in part fulfilled"); (-): rather negative ("mostly not fulfilled"); -: negative ("not fulfilled/considered"); n.a.: no assessment possible due to missing information.[12]

When it comes to the *category accuracy*, the assessments are also spread out over a large range, yet a bit further down on the scale than the other categories. Evaluations are somewhat strong in identifying and analysing the context and neutral reporting. However, they feature weaknesses in the area of the precise description of evaluation procedures including the documentation of trustworthy sources, and they do not lend themselves well to meta evaluation. The standard that was impossible to assess was the validity and reliability of the information.

On the whole, evaluations are strong or fairly strong in ensuring the identification and participation of stakeholders, in their timely reporting, cost effectiveness and in making complete and balanced assessments. However, evaluations exhibit weaknesses when it comes to the selection procedures of the evaluation team and formal written agreements, comprehensive and clear reporting, the description of the purpose and objects of the evaluation as well as evaluation procedures and methods.

Given the latter weaknesses, some evaluation reports exhibit the qualities of an expert opinion (expertise) rather than those of an external evaluation. Whereas an expert opinion is based on thematic know how, an external evaluation follows systematic if not scientific procedures (e.g., by making an assessment based on explicit criteria that are systematically applied and analysed to the empirical context).[13] The assessment team concludes that the distinctions between an expert opinion and an evaluation are not fully clear among desk

[12] In order to ensure maximum comparability, we use the same categories as Widmer 1996 (p.242).

[13] Cf. Ulrich Klöti (1997): Inhaltliche und methodische Anforderungen an wissenschaftliche Politikevaluationen, in: Werner Bussmann, Ulrich Klöti, Peter Knoepfel (eds): Einführung in die Politikevaluation, Helbling & Lichtenhahn, Basel and Frankfurt am Main, 1997, p. 39-57.

managers. Moreover, the quality assessment shows that the distinction of three categories of evaluations outlined in the evaluation program of the SDC has not taken hold. Three reports listed as external evaluations are actually external reviews according to either the desk managers and/or the title of the reports. There seems to be ambiguity on the SDC's categories of evaluations.

4.2 Strengths and Weaknesses in the TORs, the Commissioning and the Execution of the External Evaluations

4.2.1 Terms of Reference

Clearly, formulating good evaluation questions is an utmost challenge. Therefore, it is not surprising that our assessment of the TORs reveals weaknesses in this respect. For one, the statement of the primary purpose and the main objectives is often not precise enough. In addition, the purpose and objectives are not always properly linked to or consistent with the main questions listed and additional sets of sub-questions. In other cases the TORs are too complex, with various levels of questions, areas to be addressed or tasks fulfilled. What the TORs require of evaluations is thus often ambitious or not realistic given the time and financial resources allocated. One evaluation with a budget of SFR 15'000 was supposed to evaluate seven NGO projects, which it was unable to do in depth.

Unrealistic, imprecise or incomprehensible TORs are not without consequences. At best, certain questions in the TORs can only be answered in part or only superficially in methodological terms. This lowers the quality and the demand responsiveness of an evaluation. Unresolved problems in the TORs may also result in or contribute to misunderstandings among evaluators and commissioners. In the worst case these may lead to the premature termination of an evaluation project. Furthermore, problematic TORs are a hindrance for the cooperation among evaluators on an evaluation team, as they are apt to interpret the objectives of an evaluation differently among themselves.

Imprecise, inconsistent or excessively ambitious TORs appear to stem from several problems. The first is of a structural nature and a consequence of the way the TORs are drafted. The draft TORs often pass several desks (SDC desk managers and often other SDC divisions, the coordinating office abroad, possibly NGOs or other stakeholders as well as the evaluators). Thus several actors contribute to them. Not enough attention is given to consolidating the purpose, objectives and the questions asked and negotiating and agreeing on a tight final version of the TORs. A second reason for poor TORs lies in the commissioner's uncertainty as to the exact purpose of the evaluation. A third reason may be that the SDC divisions commissioning the evaluation fail to agree on the purpose and objectives. Finally, bad TORs can be due to the commissioner's lack of thematic competence and know how necessary to draft meaningful and realistic TORs.

4.2.2 Selection Procedures of Evaluation Team

The selection procedures of the sample of external evaluations got fairly negative assessments for two reasons. First, all the evaluators or evaluation teams in the assessment were appointed, in several cases only the evaluators eventually appointed were considered in the process. There were no open or near-open competitions. This also applies to those four evaluations with budgets above SFR 50'000 and those additional two evaluations which were probably budgeted above SFR 50'000 but for which we do not have the budgets for all the members of the evaluation team. It is understandable that the cost-benefit analysis of an open competition for an evaluation with a small budget of some SFR 15'000 is negative. Yet evaluations with budgets above SFR 50'000 would require a selective competition by WTO standards.

Second, the selection procedures lack transparency. Often the exact selection process cannot be reconstructed (How many candidates were discussed or through whose intervention they ended up on the shortlist, why exactly were the evaluators chosen, what criteria made the difference etc.). The lack of transparency also applies to the total cost of the evaluations, which for more than half the cases is not documented in full.

According to the desk managers the following criteria are discussed in the selection of evaluators:

- The main emphasis is on the evaluators' know how of the theme, experience in the geographical area and often also the programs, projects or institutions under scrutiny.
- Since evaluations typically rely on a participatory approach, the evaluators' confidence among stakeholders is seen as a criterion, too. Hence those responsible for the evaluandum often have a say in the selection.
- The qualifications in evaluation competence are less relevant in the selection. Desk managers take these for granted to a higher degree than thematic competences.
- The availability of evaluators at short notice is also a relevant selection criterion, yet this is not a criterion of the quality of evaluators.
- The ability to cooperate in an evaluation team was mentioned as relevant mainly in those cases where problems arose in this respect. Hence the roles and the division or labor among evaluators on an evaluation team need to be defined.

The emphasis on thematic, regional, institutional or program know how of evaluators raises the issue of *independence* of evaluators. Several desk managers alluded to this, maintaining the challenge was to find the balance between familiarity with the theme - at the risk of insufficient distance and objectivity and thus a certain "blindness" toward the evaluandum - and more distance to the theme and thus more openness – at the risk of a certain naiveté.

4.2.3 The Commissioning of the Selected External Evaluations

Most desk managers say that the commissioning starts with the definition of the needs by the desk manager in the program to be evaluated, conceding that pressure to produce evaluations may also be at work. The choice of the category of evaluation apparently receives little attention at this point. This aspect is often relegated to the E & C Division.

Insufficient awareness and/or an inconsistent use of the categories of the SDC evaluation program is a shortcoming in the commissioning of evaluations. This is also evident in that the sample of external evaluations assessed here includes three external reviews. The unclear use of categories may lead to an uneven quality of reports as well as to problems of availability of evaluation reports. The unclear use may also be responsible for formal shortcomings. This was the case with one evaluation that came close to being a self-evaluation.

4.2.4 The Execution of the Selected External Evaluations

The assessment shows several strengths of external evaluations in the execution of evaluations. First, evaluations collect information without unduly disrupting the object of evaluations. Second, they mostly ensure the participation of stakeholders and take the various interests into account. Third, they manage to adhere to the typically tight time schedules fairly well.

The record on how actively the evaluations were accompanied by the SDC is more mixed. The positive examples demonstrate regular interaction and feedback between desk managers and/or coordinating offices and the evaluators, which resulted in learning processes. The major weakness in the execution of evaluations was related to the fluctuation of personnel on the part of SDC, where job rotation is the rule.

4.3 The Quality of SDC External Evaluations Measured against Comparable Evaluations in other Swiss Government Agencies

The purpose of this section is to position the quality of SDC's external evaluations in the context of evaluations in other Swiss government agencies. This overview provides merely a global comparison of the general patterns of strengths and weaknesses. We will first draw on the findings of a study on meta evaluations of ten evaluations conducted in Switzerland[14] and then make some comparative statements on evaluation practise in Swiss government agencies in recent years.

How do the findings of the meta evaluation on evaluations conducted in Switzerland compare to this quality assessment of SDC's external evaluations? We compare the assessment by categories of standards, as the meta evaluation uses the same four groups used here - utility, feasibility, propriety and accuracy standards. The meta evaluation is based on a total of 30 standards, compared to our 23 standards. A caveat is in order on the limits of this comparison. For one, the meta evaluation is based on evaluations that were published, that had larger budgets (10 evaluations ranging from SFR 40'000 to SFR 254'000, of which 7 above SFR 100'000), longer durations (between 4 and 39 months) and were commissioned by various bodies (federal, cantonal or communal bodies and the Swiss National Science Foundation). Second, the evaluation standards used in the meta evaluation are identical neither in number nor in kind. Finally, since the sample in the meta evaluation was based exclusively on published evaluations, the availability and dissemination cannot be compared with the quality assessment here.

As far as the feasibility of evaluations is concerned, the meta evaluation study's findings are largely along the lines of those of the current quality assessment. A difference lies in the assessment of the cost effectiveness, which is higher in SDC's external evaluations. However, this difference can be explained with the much higher average budget of the evaluations in the meta evaluation. Higher budgets would seem to go hand in hand with higher expectations.

The meta evaluation draws a generally positive assessment on propriety standards. Our conclusions on SDC's external evaluations are similar on standards such as complete and balanced assessment and describing conflicts of interests. Yet SDC evaluations are rated somewhat more negatively on formal written agreements. The quality assessment found gaps or unrealistic expectations in these agreements which harbor the potential for misunderstandings.

As for the utility standards, the meta evaluation has positive ratings on the competence of evaluators, the transparency of value judgments and the timeliness of evaluations. SDC evaluations also tend to be fairly strong on these standards. But they are clearly stronger than those in the sample of the meta evaluation when it comes to the identification of stakeholders. SDC's strength here is in line with its emphasis on close cooperation with stakeholders which desk managers value highly. Nevertheless, SDC's external evaluations are weaker regarding comprehensive and clear reporting. This seems to be due to two reasons. First, SDC's evaluations have much smaller average budgets so that lower expectations on comprehensive reports seem to be assumed. Second, the sample in the meta evaluation consists of published evaluations, whereas SDC's evaluations seem to define utility more narrowly since its reports are not as accessible to outsiders.

The meta evaluation identifies the most glaring weaknesses in the accuracy standards of evaluations. The sample reveals weaknesses similar to those of SDC's evaluations: The

[14] Widmer, Thomas (1996): Meta-Evaluation, Kriterien zur Bewertung von Evaluationen, Bern, Haupt.

clear description of the objectives of evaluations and the description of evaluation procedures and methods are insufficient. Interestingly, the meta evaluation finds this shortcoming to apply in particular to qualitative evaluations. SDC's external evaluations are typically qualitative in nature and thus fit this pattern.

If we place SDC's external evaluations in the context of today's *evaluation practise in Swiss government agencies* based on our personal evaluation experience, we come to the following conclusions:

- User orientation and the acceptance of evaluations and recommendations among desk managers are high in SDC's external evaluations. This seems to be a strength compared to other government agencies.
- Selection procedures tend to be more transparent in other government agencies than they are for SDC's external evaluations.
- There are generally too many questions asked in the TORs of evaluations in government agencies. Yet the SDC's weakness appears to be more pronounced than in other agencies.
- The availability of SDC's external evaluation results is weaker on average than this is the case for evaluations of other government agencies.

4.4 The Levels and the Objects of SDC's External Evaluations

All evaluations focus on the output *levels*.[15] This is not surprising since output issues are more easily addressed so that evaluations generally speaking tend to exhibit the best quality at this level. A large majority of SDC's external evaluations also requires assessments at the outcome and impact levels, which means that evaluations often combine assessments at all three levels. The levels are at times explicitly and clearly mentioned in the TORs, at times they become clear only indirectly from the description of tasks. In some cases there is no explicit reference to levels, or they are mixed up and unclear.

In theory the *targeted levels* are in many cases *appropriate for the object of evaluation*, including impact levels in those cases where a program or project has been in effect long enough to make an assessment meaningful. The object of evaluation thus may often warrant the consideration of two or three levels – in theory.

The question whether the targeted level is *appropriate measured against best practises* needs to be answered in relation to the resources and duration of the evaluations. The focus on the output level is in general appropriate and realistic. For projects with as yet short life spans, the outcome level is difficult or impossible, the impact level impossible to evaluate. For longer-running projects a focus on the outcome and impact levels is appropriate as long as the evaluation projects provide the necessary resources and duration. However, within the short time frame and within the modest budgets, it was often unrealistic to expect systematic analyses of the impact of programs or projects. The expectations regarding evaluation of outcome and impact aspects were often too ambitious and not appropriate against best practise. Some reports conceded the limited nature of the assessment on outcome and impact levels.

Generally, the *object of evaluation* at hand is often appropriate in theory, which means it would lend itself to evaluation in principle. However, when seen in relation to the budget and time frame available, the evaluations cannot fulfill expectations. In several evaluations the

[15] The *output level of evaluation* attempts to assess services provided by a project or program. The *outcome level* assesses a project or program's influence on the immediate target groups. The *impact level* assesses the socio-economic and socio-political effects.

object is too broad – such as the number of projects to be evaluated too large - when considered against resources allocated.

4.5 The Cost Effectiveness of SDC's External Evaluations

The cost effectiveness receives high marks in this assessment: SDC's external evaluations generally produce information of a value that justifies the cost of producing them. In short, SDC gets value for money. However, we concede that the assessment may be too positive for two reasons. First, the small budgets in absolute terms may lead to a more positive view easily. Second, the judgment is based to a large degree on the desk managers' views and is therefore skewed. What we criticise in the context of the cost of external evaluations is the insufficient transparency. The total cost is documented only for seven evaluations.

4.6 Conclusions Regarding the Use of the Evaluation Results in Decision-making

In terms of demand responsiveness the evaluations were assessed as middle range to fairly positive. Recommendations were generally seen as useful and realistic for decision-making by desk managers, regardless of whether they were in fact implemented or not. Yet we need to consider the bias of this assessment because only desk managers were interviewed.

With other target groups in mind, the use of evaluation results in decision-making might be judged more critically. We were therefore critical in the assessment of the availability and the accessibility (in the sense of comprehensibility) of evaluations for actors outside the immediate group of stakeholders ("non-insiders"). Accessibility also needs to be seen in the light of accountability. We believe that the results of external evaluations should be made available to a broader audience:

- "New" desk managers: Given personnel fluctuation and staff rotation in SDC and the need to ensure continuity in accompanying and utilising evaluations, evaluation reports should be more accessible.
- Other SDC divisions: Some evaluations would lend themselves to utilisation in other SDC divisions, such as an evaluation of a development fund, partnerships or return programs. The dissemination and discussion of results (good practices, lessons learnt, recommendations, benchmarks) should thus be broadened to a wider circle where this is the case.
- SDC (senior) management: The wider the potential consequences of an evaluation, the more an evaluation needs to be anchored broadly within SDC. This may mean higher up in management or in more than one SDC division.
- Politicians: Even if political actors are not the primary target audience of SDC external evaluations, politicians may still be interested in some evaluation results.
- Partner organisations such as development agencies in other countries, international organisations such as the OECD as well as NGOs.

5 Recommendations

In this chapter we present our recommendations aimed at improving the use of evaluation results and moving in the direction of more focused, more standardised and higher quality external evaluations over the short and the long term.

5.1 Draft Good, Realistic and Comprehensible TORs

First and foremost, the SDC needs to have better TORs. In order for the SDC to reach this goal, we recommend that those commissioning evaluations follow these steps:

- Formulate a clear purpose to an evaluation. Identify the users of the evaluation and involve them in the process of drafting the TORs.

- Define objectives that are clear, focused, concise and understandable. Limit the number of main questions to three or four.

- Ensure agreement and acceptance among SDC divisions if more than one division has a stake in the evaluation.

- Be realistic about what an evaluation can do. Consult experienced evaluators on whether the objectives match the resources allocated. Pay attention to these aspects:

    - Restrict the focus of the evaluation to one or two levels of evaluation (e.g., output and outcome levels).[16] Beware of questions at the impact level, which requires both a program in effect long enough for effects to ensue and more time and resources to study them.
    - Appreciate the time it takes to write good reports when fixing the schedule.

- Before the TORs are finalised, have them checked and consolidated by those responsible for evaluation in the departments.

5.2 Ensure More Competitive and Open Selection Procedures for Evaluation Teams

Striking a balance in the evaluator's profile between the desired thematic and regional know how and ensuring a degree of independence from the project at hand is a daunting task. Thematic know how involves closeness with and a potential "blindness" to the project and its stakeholders, whereas a higher degree of independence means more distance if not "naiveté" in view of the project. We recommend that the E & C Net develop guidelines considering the following aspects of the selection process:

- Provide more openness and transparency in the selection procedures in general. Open competitions provide an environment conducive to innovation. Be prepared to learn from less preconceived, more "naive" voices.

- Place more emphasis on evaluators' experience in evaluation (skills in methods).

- Leave the team building to the lead evaluator (if there is an evaluation team), as it requires attention to matching personalities and skills. Nevertheless, ensure that the division of labor and responsibilities among the team be defined (formal agreement).

- Encourage the participation of local consultants in the evaluation teams where this is warranted by the object of evaluation.

5.3 Improve the Conditions for the Utilisation of External Evaluations

We suggest that the desk managers improve the conditions for the utilisation of external evaluations in the following ways:

- Clarify the roles of the actors commissioning and/or using an evaluation at the outset, such as the SDC divisions involved (if more than one), the desk manager, the Coordinating Office abroad and other actors.

[16] The three levels are explained in 3.1.2.

- Accompany evaluations closely. Continuous contact between users and evaluators is key and enhances the learning potential on the part of the SDC. In the event of staff rotation, ensure a smooth transition between desk managers by requiring a formal hand-over of running evaluations.

- Do not accept external evaluation reports written exclusively for insiders and stakeholders narrowly speaking. Both structure and content of the reports need to be conducive to understanding for a wider audience. The E & C Net needs to support the desk managers commissioning evaluations in the quality management (e.g. by providing feedback on the quality of reports). Enhancing the quality of evaluation reports also facilitates staff rotation.

- Make external evaluation results more widely available. For this purpose the E & C Net needs to maintain and update data banks so that the results may be accessed with ease. An enlarged target audience[17] as well as an increase in accountability will ensue.

5.4 Enhance the Conditions for High Quality External Evaluations

In an effort to enhance the quality of external evaluations, desk managers need to get support in various respects. We recommend that the E & C Net launch a coordinated effort to establish support structures including steps such as the following:

- Provide a systematic training scheme in evaluation addressing questions such as: What is evaluation and how does it differ from expert opinions? What needs to be considered in defining the object of evaluation (e.g. levels of evaluation)? How do the three categories of the SDC evaluation program differ and what are their corresponding requirements? What procedures do evaluations involve? What are the success factors for the utilisation of evaluations?

- Provide support in drafting the TORs and commissioning evaluations. Give guidance in selecting the appropriate category of evaluation and dealing with the corresponding requirements. Emphasise the importance of realistic time schedules and mission plans as well as sufficient time for writing the reports.

- Provide support in accompanying evaluations. This includes support in addressing problems of cooperation among the members of the evaluation team as well as constraints encountered by the evaluators.

- Provide support in obtaining high quality evaluation reports by emphasising feedback for desk managers. The feedback needs to address such issues: Are the purpose, the main objectives and the key questions clearly described? Are evaluation procedures and methods described sufficiently? Is the structure of the report logical? Are results, conclusions and recommendations presented in a coherent and substantiated fashion?

To sum it up, “less may be more”. In other words, the SDC conducts too many evaluations with too little resources. Realistic expectations need to translate into more resources per external evaluation. If resources for evaluation remain stable, the consequence should be fewer, but well-prepared and well-implemented evaluations.

[17] See 4.6 for an elaboration of an enlarged target audience.

6 Comment on the Use and Applicability of the Evaluation Standards

What feedback can be provided on the utility and applicability of the evaluation standards used in this quality assessment? In chapter 2.2 we outlined the steps used to produce the list of 23 evaluation standards applied here. These standards are based on the DAC Minimum Sufficient Evaluation Standards, the SEVAL standards of the Swiss Evaluation Society and the key questions of the Approach Paper. The list of criteria in the appendix indicates the corresponding source(s) of each standard.

In setting up the list of standards, we attempted to minimise the overlap among the evaluation standards as far as possible. However, a certain degree of overlap is inherent in standards and categories. We divided the standards into the four categories of standards according to the SEVAL (utility, feasibility, propriety and accuracy) rather than the six from the DAC standards (evaluation purpose, design and implementation of evaluations, credibility, usefulness, impartiality and independence, reporting). In our view the SEVAL categories are more focused and to the point, more manageable, easier to delineate and thus easier to apply for the purposes of a quality assessment. The standards that we found missing in the DAC standards yet necessary for a quality assessment are presented below with the category indicated in parenthesis:

- *Making findings available (utility):* Results of an evaluation need to be made available to all potentially affected people as well as those with a legitimate claim to receiving the results (SEVAL standard P5). We believe this standard is important enough that it warrants a standard of its own.

- *Evaluation impact (utility):* SEVAL standard U8 „Evaluation Impact“ states that the planning, execution and presentation of an evaluation should encourage stakeholders both to follow the evaluation process and to use the evaluation. This standard contributes to the utility/usefulness of an evaluation.

- *Cost effectiveness (feasibility):* The question whether evaluations produce information of a value that justifies the cost of producing them is both a SEVAL standard (SEVAL standard F3) and a key question in the Approach Paper.

- *Trustworthy sources of information (propriety):* The DAC standards in the section „Design and Implementation of Evaluations“ do not address this issue sufficiently (6 and 7) (SEVAL standard A4).

- *Analysing the context (accuracy):* Identifying the influence of the context on the object of evaluation seems particularly important in areas as diverse thematically and geographically as they are in the field of development and cooperation. Evaluations should thus include relevant information on issues such as institutional embeddedness, the social and political climate, the characteristics of and relationships among the key stakeholders (including the commissioning party), possible hidden agendas etc. (SEVAL standard A2 „Analyzing the Context“ and F2 „Anticipating Political Viability“).

- *Meta evaluation (accuracy):* The standards should include mention of necessary meta evaluation (or „quality assessments“ in SDC parlance) so that stakeholders can assess the evaluation’s strengths and weaknesses, which enhances the utilisation of evaluation. (SEVAL standard A10)

On the whole the standards proved to be fairly good and applicable in this quality assessment. Nevertheless, there were some problems due to the complexity of a few standards on the one hand and overlap among some standards on the other. On account of

the complexity of standards in the context of this fairly small-scale quality assessment it was difficult to assess the following standards:

- demonstrated professionalism and competence (4),
- cost effectiveness (12), and
- valid and reliable information (20).

For the assessment of the first two of these standards we relied largely on the desk managers' views. The latter standard was impossible to assess as the external evaluations in the sample largely failed to provide transparent and clear descriptions of evaluation procedures and methods. Despite these shortcomings related to the complexity of standards, we deem it important for a quality assessment to attempt to say something about the professionalism of the evaluators as well as the cost effectiveness of an evaluation. However, the standard valid and reliable information seems to be of any worth only if there are distinct improvements in the description of evaluation procedures and methods. Based on our assessment as well as the meta evaluation study of 1996,[18] qualitative evaluations are especially weak in this area.

The problem of overlap arose particularly between the following standards from our list of evaluation standards. In the first case, we suggest how to deal with the problem (in italics).

- Overlap between standard 9 evaluation impact and standard 11 evidence of participation: *Delete the last question of standard 9 ("Are the various positions involved taken into account in order to win their cooperation?").*

- Overlap between standard 2 evaluation purpose and objectives clear and standard 13 formal written agreement.

- Overlap between standard 14 complete and balanced assessment and standard 22 neutral reporting.

- Overlap between standard 7 transparency of value judgments and standard 21 impartial and substantiated conclusions.

On the whole, we consider the problems of overlap to be minor as well as inevitable in an endeavor such as a quality assessment of evaluations. We thus suggest that the standards be maintained and an effort to minimise the overlap be made by cross-referencing standards where appropriate.

[18] Widmer, 1996.

Institut für Politikstudien

I N T E R F A C E

Seidenhofstr. 12
CH-6003 Luzern
Telefon 041 412 07 12
Fax 041 410 51 82
www.interface-politikstudien.ch

Appendices

List of Evaluation Standards Used in the Quality Assessment

No.	Standard (DAC, SEVAL)	Questions	Comment	Assessment					
				+	(+)	+/–	(–)	–	k.A
Utility									
1	*Stakeholders identified* (DAC 2, SEVAL U1)	Does the report identify (I) the ultimate beneficiaries of the evaluation, (II) the core learning partners, (III) those best positioned to implement the recommendations of the evaluation?							
2	*Evaluation purpose and objectives clear* (SEVAL A1, U2, DAC 1, DAC 3, key question 4 of SDC Approach Paper)	Do the Terms of Reference and the report clearly state the primary purpose and the objectives of the evaluation? Is the process adopted to ensure that all stakeholders understand the objectives described? Is the level of evaluation (output, outcome, impact etc.) clearly stated and appropriate?							
3	*Demand responsive* (DAC 11, DAC 14, SEVAL U4, Approach Paper key question 7)	Is the evaluation focused on the central questions of the TORs and does it answer them all and in a way that reflects their stated level of priority? Are the recommendations useful? Is the object of evaluation appropriate?							
4	*Demonstrated professionalism and competence* (SEVAL U3, DAC 9)	Can the evaluation team be considered as credible regarding its qualifications and experience in the evaluandum?							
5	*Selection procedure of evaluation team*	Was the evaluation team appointed directly or was there a competition? Were those responsible for the evaluandum able to influence the choice?							
6	*Comprehensive and clear reporting* (SEVAL U6, DAC 5, DAC 20, DAC 13)	Does the evaluation report precisely describe the object of evaluation? Is the evaluation report logically structured and does it outline the evaluation context, goals, questions posed, and procedures used, as well as any constraints encountered that substantively hindered its ability to fulfill its purpose? Is there an executive summary with key findings, conclusions and recommendations?							
7	*Transparency of value judgments*	Are the underlying reasoning and points of view upon which							

No.	Standard (DAC, SEVAL)	Questions	Comment	Assessment					
				+	(+)	+/–	(–)	–	k.A
	(SEVAL U5, DAC 17) Answer DAC 17 under standard 21)	an interpretation of evaluation results rests described in such a manner that the bases for the value judgments are clear?							
8	*Timely reporting* (SEVAL U7; last sentence DAC paragraph on “Usefulness”)	Have significant interim results as well as the final report been made available to the intended users in such a way that they can be utilized in a timely manner?							
9	*Evaluation impact* (SEVAL U8, DAC 19) answer DAC 19 under standard 11	Do the planning, execution, and presentation of the evaluation encourage stakeholders both to follow the evaluation process and to use the evaluation? Are the various positions of the different interests involved taken into account in order to win their cooperation?							
Feasibility									
10	*Practical procedures* (SEVAL F1)	Are evaluation procedures designed in a way that the information needed is collected without unduly disrupting the object of the evaluation?							
11	*Evidence of participation* (SEVAL F2, DAC 19)	Did the stakeholders have the chance to participate and introduce their views?							
12	*Costs and cost effectiveness* (SEVAL F3, DAC 12)	What were the costs of the evaluation? Does the evaluation produce information of a value that justifies the cost of producing them (value for money)?							
Propriety									
13	*Formal written agreement* (SEVAL P1, DAC 4)	Is there a formal written agreement specifying the duties of the parties who agree to conduct an evaluation? Does this agreement clearly state the areas to be addressed by the evaluation [scope of work], the key questions, the resources and the time allocated, methodology and procedures to be followed, and reporting requirements?							
14	*Complete and balanced assessment* (DAC 16, SEVAL P4)	Is the evaluation complete and balanced in presenting and assessing the strengths and weaknesses of the object being evaluated?							
15	*Making findings available* (SEVAL P5, feedback 2.1)	Are the results made available to all the potentially affected persons as well as to all others who have a legitimate claim to receiving them?							
16	*Declaring conflicts of interest* (SEVAL P6, DAC 15)	Are conflicts of interest addressed openly and honestly so that they compromise the evaluation process and conclusions as little as possible?							

No.	Standard (DAC, SEVAL)	Questions	Comment	Assessment					
				+	(+)	+/–	(–)	–	k.A
Accuracy									
17	*Identifying and analysing the context* (SEVAL A2, DAC 8 last sentence)	Are the influences of the context on the object of evaluation identified and described?							
18	*Precise description of evaluation procedures* (SEVAL A3, DAC 6, DAC 10)	Does the evaluation report contain a detailed description of the organisation of the evaluation, data collection and processing, analysis and reporting? Are the procedures used sufficiently precisely described and documented so that they can be identified as well as assessed? Is the choice of method discussed in the report?							
19	*Trustworthy sources of information* (SEVAL A4)	Are the sources of information used in the evaluation sufficiently precisely described so that their adequacy can be assessed?							
20	*Valid and reliable information* (SEVAL A5, A6, A7, DAC 7, DAC 8)	Are the data collection instruments selected, developed and employed valid and reliable? *Validity* is determined by assessing the degree to which the instruments employed accurately reflect the concepts they are intended to measure. *Reliability* refers to the consistency or stability of the quality measured, whether between measurement instruments, persons, or over time. Are methods and procedures applied as stated and in accordance with their own quality standards (e.g. statistical tests, validity thresholds, attrition biases)?							
21	*Impartial and substantiated conclusions* (SEVAL A8, DAC 17. Restriction: our feedback 1.2 on DAC 17)	Are the conclusions reached in the evaluation clearly and explicitly described and substantiated in such a manner that stakeholders can comprehend and judge them?							
22	*Neutral reporting* (SEVAL A9, DAC 18)	Is the evaluation report free from distortion through personal feelings or preferences on the part of any party to the evaluation? Does the evaluation report present conclusions in a neutral manner?							
23	*Enabling metaevaluation* (SEVAL A10)	Is the empirical material of the evaluation available in order to enable a metaevaluation to check if the evaluation is appropriately executed so that stakeholders can assess the evaluation's strengths and weaknesses?							

Institut für Politikstudien

Seidenhofstr. 12
CH-6003 Luzern
Telefon 041 412 07 12
Fax 041 410 51 82
www.interface-politikstudien.ch

DAC Minimum Sufficient Evaluation Standards (DAC Standards)

Evaluation Purpose

The main purposes for the conduct of evaluations are:
to improve future aid policy, programmes and projects through feedback of lessons learned;
to provide a basis for accountability, including the provision of information to the public.[19]

1. **Clear Purpose** The Terms of Reference and the Evaluation Report clearly state the primary purpose of the evaluation as determined by the commissioning body.

Design and Implementation of Evaluations
Each evaluation must be planned and terms of reference drawn up which adequately define the purpose, the evaluation issues to be addressed, stakeholders, methodology, performance standards, resources and budget required to complete the evaluation.[20]

2. **Stakeholders Identified** The report clearly identifies the stakeholders participating in, and affected by, the evaluation. Specifically, the report identifies: (i) the ultimate beneficiaries of the objective of evaluation; (ii) those persons who most need to learn from the evaluation; and (iii) those who are best positioned to implement the recommendations contained in the evaluation report.

3. **Evaluation Objectives Clear** The report clearly describes the objectives of the evaluation and the process adopted to ensure that all stakeholders understand the objectives of the evaluation. Clarifying the objectives of an evaluation is often not fully possible at the outset of an evaluation, but instead calls for a lengthier process that should be regarded as a central element of the evaluation process itself. Where this is so, the process of clarifying the objectives is clearly described.

4. **Relevant Scope** The Terms of Reference [TOR] and the Evaluation Report clearly state the areas to be addressed by the evaluation [scope of work]; the information identified for collection; the standard against which performance is to be assessed or analyses are to be conducted; the resources and time allocated and reporting requirements. The TOR render it possible to ask pertinent questions about the object of evaluation and take into account the interests and needs of the parties commissioning the evaluation, as well as other stakeholders.

5. **Precise description of the Object of Evaluation** The object of evaluation, be it a measure, program, or organization, is clearly and precisely described, documented, and unambiguously identified. Particular attention has been paid to any discrepancies between the original form the object of evaluation was anticipated to take and its actual form in practice or when implemented.

6. **Defensible Methodology** The questions to be addressed in the evaluation, and the methods and procedures chosen to address these questions, have been carefully

[19] Development Assistance Committee "Principles for Evaluation of Development Assistance" OECD Paris 1991 Section II paragraph 6.
[20] Ibid Section IX paragraphs 32 ff

documented. Contextual constraints have been identified and methods for dealing with these constraints have been explained. The report contains a detailed description of the organization of the evaluation, data collection and processing, analysis and reporting. Major methodological options have been discussed, including the risks associated with alternative options, and choices justified. Selected methods and procedures have been applied as stated and in accordance with their own quality standards (e.g. statistical tests, validity thresholds, attrition biases). Limitations faced in data collection and analysis have been described. Any changes that have occurred in proposed methods and procedures during the course of the evaluation have been described and justified.

7. **Valid and Reliable Information** The data collection instruments selected, developed and employed are valid and reliable. Validity is determined by assessing the degree to which the instruments employed accurately reflect the concepts they are intended to measure. Reliability refers to the consistency or stability of the quality measured, whether between measurement instruments, persons, or over time. All potential biases or errors are systematically identified, analysed and corrected as far as possible by recognised techniques.

8. **Sound Analysis** Data are appropriately and systematically analysed or interpreted according to the state of the art. Major cause-and-effects relationships and underlying assumptions are made explicit. Critical exogenous factors have been identified and taken into account.

Credibility

The credibility of an evaluation depends on the expertise and independence of the evaluators and the degree of transparency of the evaluation process.[21]

9. **Demonstrated Professionalism and Competence** The evaluation report demonstrates the competence and trustworthiness of the evaluators [Note: Performance with respect to this standard must be deduced rather than directly assessed. Evaluation reports assessed as being of a high quality with respect to other standards (e.g. 8-20) would be rated high with respect to this standard].

10. **Transparent Evaluation Process** The evaluation report contains a clear and sufficient explanation of the process and methods for conduct of the evaluation that is accessible to relevant stakeholders [Note: Performance with respect to this standard is partially dependent upon assessed performance with respect to 12, 13, 14, 17, 20].

Usefulness

For an evaluation "to have an impact on decision-making, the findings, conclusions and recommendations must be perceived as being relevant and useful and be presented in a clear and concise way. They should fully reflect the different interests and needs of the many parties involved in development cooperation. Easy accessibility is also crucial for usefulness. The evaluation process itself promotes a further clarification of objectives, improves communication, increases learning, and lays the groundwork for follow-up actions. Evaluations must be timely in the sense that they should be available at a time that is appropriate for the decision-making process."[22]

[21] Opcit OECD Paris 1991 Section IV paragraph 18.
[22] Ibid Section V paragraphs 21 and 22.

11. **Demand Responsive** The evaluation report adequately addresses the information needs of the commissioning body. It answers all question included in the Terms of Reference in a way that reflects the stated level of priority.

12. **Robust Findings** The report provides stakeholders with a substantial amount of new knowledge (findings). Findings are clearly identified. They follow logically from, and are justified by, data, interpretations and analyses through logical reasoning that are carefully described and do not contradict each other. Logical reasoning is developed as far as possible and necessary. When relevant, the report indicates which findings are generalisable and under which conditions.

13. **Clear Conclusions** The conclusions reached in the evaluation report are clearly and explicitly described, together with their underlying assumptions.

14. **Useful Recommendations** Recommendations are not mixed with conclusions, but they are derived from them. Recommendations are presented in sufficient detail and with an operational focus. The report indicates that practical constraints have been taken into account when formulating recommendations (e.g. regulations, institutions, budget).

Impartiality and Independence

"Impartiality and independence are closely inter-related concepts. In fact, the aim of impartiality is best achieved where evaluation activities are independent from operations personnel and managers who have interests in showing accomplishments and good performance. Impartiality also depends on the professionalism of evaluators and the methodology applied."[23]

15. **Impartial and Independent Evaluation Function** The evaluation report clearly indicates the degree of independence of the evaluation function from the operations and management functions. Conflicts of interest are addressed openly and honestly so that they compromise the evaluation process and conclusions as little as possible.

16. **Complete and Balanced Assessment** The evaluation report is complete and balanced and presents the strengths and weaknesses that exist in the object being evaluated, in a manner that strengths can be built upon and problem areas addressed.

17. **Impartial and Substantiated Conclusions** The process employed in reaching conclusions is described. [It should be noted that conclusions go further than findings in the sense that they include value judgements.]
The conclusions are based on explicit and agreed judgement criteria and benchmarks. The judgement criteria take into account all legitimate standpoints. Conflicting points of view and issues are presented in a balanced way. There are no discrepancies between stated criteria and benchmarks and those that have been actually applied.

18. Neutral Reporting The evaluation report is free from distortion through personal feelings or preferences on the part of any party to the evaluation. Evaluation reports present conclusions in a neutral manner.

Participation of Donors and Recipients

[23] DAC Working Party on Aid Evaluation "Review of the DAC Principles for Evaluation of Development Assistance OECD Paris 1998 Pg 24

"Whenever possible both donors and recipients should be involved in the evaluation process. „[24]

19. **Evidence of Participation** The evaluation report details the way in which donor/recipient participation has been encouraged in the planning, execution and presentation of the evaluation. Where the purpose of an evaluation is to investigate the impact of the object of evaluation, be it a measure, program or organization, on the lives and welfare of beneficiaries evidence of participation/consultation with those beneficiaries is provided.

Reporting

20. **Comprehensive and Clear Reporting** The final evaluation report is logically structured and outlines the evaluation context, goals, questions posed, and procedures used, as well as any constraints encountered that substantively hindered its ability to fulfil its purpose and adhere to good evaluation practice. The findings, conclusions and recommendations reached are outlined in such a manner that the most pertinent information is readily accessible and comprehensible.

The report is free of superfluous information and analyses that do not substantiate the conclusions.

The report contains a short executive summary that highlights the key findings, conclusions and recommendations in a balanced and impartial way. Only appendices contain technically difficult information that is not accessible to all stakeholders.

[24] Ibid Section VI paragraph 23