To evaluate in a credible and meaningful way: between dream and reality

A study of the evaluability of (co)-financed interventions of the Belgian Cooperation
Study of the Evaluability of (co-) financed interventions of the Belgian Cooperation

Nathalie HOLVOET - Liesbeth INBERG - Bob PEETERS - Lisa POPELIER - Dirk VAN ESBROECK - Ellen VERHOFSTADT

Final Report

February 2016

This Study was executed by a partnership between South Research and IOB (University of Antwerp), and supported by a steering committee.

The opinions expressed in this document represent the evidenced view of the authors and do not necessarily reflect the position of the FPS Foreign Affairs, Foreign Trade and Development Cooperation.
© FPS Foreign Affairs, Foreign Trade and Development Cooperation

February 2016

Graphic design: Sandra Fernandez Bernardo, FPS Communications Unit

Printed by: FPS Printing Office

Evaluation no. S4/2014/03

Legal Deposit: 0218/2016/012

This document is also available on the website http://diplomatie.belgium.be/en/policy/development_cooperation/how_we_work/special_evaluation_office and at the Special Evaluation Office for the Belgian Development Cooperation.

This report should be cited as follows:
Special Evaluation Office of the Belgian Development Cooperation/ SEO (2016), To evaluate in a credible and meaningful way: between dream and reality, a study of the evaluability of (co)-financed interventions of the Belgian Cooperation, FPS Foreign Affairs, Foreign Trade and Development Cooperation, Brussels.
Executive Summary

Background of the study

This study has been carried out against the background of the increasing importance of evaluations in development cooperation on the one hand, and the perception that until now, the quality, usability and effective use of evaluations for several reasons have been under par, on the other. The first phases of the study were carried out by SEO and formed an important base for the second phase. According to the assignment specifications the general purpose of the study was to ‘contribute to making all future actions evaluable in the not too long term’.

The study comprised a total of 40 interventions in 4 countries (Belgium, Benin, DRC and Rwanda). The interventions were carried out in several sectors via a wide range of development stakeholders and were (co)financed by the Belgian federal government. Using a study framework (see below) the interventions were given a score for several elements that affect evaluability. The resulting scores enable comparison of the evaluability of the 40 interventions, but do not allow for general conclusions with respect to the degree of evaluability of the interventions of the Belgian development cooperation in general.

Basic concepts

The relative unfamiliarity with the concept of evaluability and the initial reservations some stakeholders had with respect to this study, prompted the study team to start by building broad support for it and providing unambiguous definitions of the concepts evaluability and evaluability assessment. The study uses the OECD/DAC definition which defines evaluability as “the extent to which an activity or a program can be evaluated in a reliable and credible fashion”. Evaluability assessment is often associated with evaluability and described as “an instrument which allows to assess whether a certain situation calls for an evaluation”.

Another important distinction is made between theoretical evaluability and practical evaluability. Theoretical evaluability refers to evaluability in principle as it can be derived from the design of the intervention, without taking practice into the equation. Practical evaluability does take practice into account and will check, for instance, whether the data concerning the progress of an intervention is, in fact, collected and whether M&E systems are actually being used, so that they can contribute to improving management and results.

The study framework we developed took account of these concepts and was broken down into three parts, called ‘dimensions’ in the study: the intervention design; the practice with respect to implementation and management of the intervention; and the role of contextual factors. The three dimensions were further divided into eight components, which in turn were divided into sixty-two items. This study framework is an important result of the study and a useful instrument for anyone who wants to do more in-depth research into the subject of evaluability.

This study approaches evaluability mainly as a continuum. According to this approach, no intervention is perfectly evaluable in all its aspects, while on the other hand it is quite unlikely that there are interventions that cannot be evaluated at all. Furthermore, an
evaluation may comprise several aspects of an intervention – the five classical OECD/DAC or other criteria; it may have several purposes (focus on accountability and/or learning, etc.); and it may or may not be controlled from an evaluation policy. Evaluability will be different in each of these situations.

**The importance of “evaluability” and “evaluability assessment”**

Determining the evaluability of an intervention does not serve to determine the (development) value of that intervention. It is perfectly possible that highly valuable interventions are hard to evaluate. In other words: there is no connection between (the extent of) evaluability and the (development) value of an intervention. On the other hand, the importance of (the concept of) evaluability and (the instrument) evaluability assessment is undisputed, despite their relatively limited use. They may, for instance improve the quality of evaluations, and ensure that the evaluation results are put to more effective use. A good evaluability assessment can be carried out at a fraction of the cost of the entire evaluation budget, but it can make a huge difference. Furthermore – as will be confirmed by the results of this study – the field of application of an evaluability assessment is not limited to evaluation as such; assessment may in all of the phases of an intervention contribute to better management.

**Key findings and conclusions**

**Knowledge and use of evaluability and evaluability assessment** The study found that until now, the concept evaluability and the instrument evaluability assessment (or evaluability test) are little known and used in the Belgian development cooperation environment. Elements of evaluability tests are practiced here and there – without naming them as such – but there is no systematic use. This finding is of interest as in recent decades, evaluation has become increasingly important and has grown into a fixed component of management practices in the development cooperation environment: today stakeholders actually cannot afford not to evaluate. However positive this development may be, it also holds the risk that evaluations turn into ritual exercises without any true involvement of the key stakeholders. A more conscious use of the concept of evaluability and of the instrument evaluability test, including the option of coming to well-founded statements about the desirability of an evaluation, may in this context prove to be an important instrument towards improved and more relevant implementation of the role and function of evaluation(s) in the environment of development cooperation.

**General findings.** The 40 analysed interventions returned an average score (based on the 62 items in the study framework) of just above the centre point of the scale. Although we cannot attach an absolute meaning to the scores, they give a good indication of the most important general finding from this study: interventions have a number of strong points, but they require work in terms of improving their evaluability. Furthermore, the distribution of the scores approximates a normal distribution, but its large spread implies that despite controlling action from the funding government, the Belgian stakeholders and intervention types show considerable differences in terms of quality of management.

**Stronger and weaker components and their mutual influencing.** Of the three studied dimensions (intervention design, implementation practice and context), the latter scored considerably higher than the former two. This illustrates, again, the margin for improvement of the dimensions that stakeholders have most control of (intervention design and implementation practice). More specifically, the lowest scores were reported for the components the proposed M&E system; the quality of the basic information with respect to the intervention implementation; and the quality of the intervention logic and theory of change. The highest scores were found for the components distinguished under the context: the attitude of the key stakeholders towards (external) evaluation and the influence of the (political and) institutional context. Furthermore, it has become clear
that irregularities in the intervention design often have consequences for the implementation and thus affect evaluability both directly and indirectly. After all: a good design phase is often the first indication of high quality intervention management. Investing in proper intervention preparation appears to pay itself back, whereas initial weaknesses are harder to correct at a later stage than one would expect on the outset; in other words: the initial design turns out to be the reference on which practice is based.

**Several consequences of weaknesses in the intervention design for monitoring and evaluation.** It appears that a weak design phase has different consequences for monitoring and evaluation. Apparently it is not too hard to apply corrections at the – mainly operational – level of monitoring. This shows from the fact that the quality of the M&E system in practice is considerably higher than the quality of the proposed M&E system, although hiatuses from the initial phase do have a continuous effect on the implementation. However, hiatuses in the intervention design have more severe consequences for the evaluation function because corrections are harder to apply here and they imply that certain important aspects of an intervention, such as effectively reaching initial target groups, the effects of the intervention on several social groups, and the realisation of assumptions and risks, cannot or hardly be evaluated and often are excluded from evaluations altogether. In that case, the combination of the aforementioned factors can result in ‘independent’ evaluations being controlled (or at least determined) largely from within, which ultimately means that gaps and blanks deliberately or accidentally will not be acknowledged and controversial subjects will be avoided or receive too little attention. In other words, there is a risk that evaluations focus only on reality as it has been defined or interpreted by the intervention involved.

**The general evaluability scores per (OECD/DAC) evaluation criterion** yield lower scores for sustainability and particularly impact across all the dimensions and components. The fact that the same score pattern keeps returning obviously has to do with the degree of difficulty – at all stages of the intervention cycle – of the evaluation of these criteria. Although generalisation is hard in this respect, it can be said that sustainability and more particularly, impact are harder to evaluate than the three other criteria. Impact evaluations require a high level of methodology (and funds), while the difficulty of sustainability evaluations often lies in the challenge of making well-founded statements about a situation that will happen no sooner than in the future. The study also found that despite increased attention for sustainability, it is insufficiently integrated in the management systems. On the other hand, effectiveness and efficiency show considerably higher scores – an indication of the quality of the intervention management, particularly as regards monitoring and evaluation. In this context it is important to recognise the influence of the funding government. DGD’s main concern is the correct use of the public funds that have been made available and for that purpose (by way of various procedures and formats, etc.) has imposed important conditions for the management of funded interventions with a clear focus on efficiency, whereas an aspect such as impact seems less prominent. Furthermore, the high evaluability score for efficiency can be explained from the important efforts made by BTC and NGOs in particular to further develop and implement their M&E systems. These efforts are the result of processes that have been applied within these organisations for some time, but that are also a reaction to the planned screening of NGAs in 2016 which will be examining the quality of the M&E systems. It should be noted that the development of M&E systems often involves rather complex initiatives that evolve gradually, with the changes at intervention level taking place rather slowly. These processes follow a bottom-up approach in which the implementation level (inputs – activities – outputs) comes first; incidentally, this is also the level at which the direct usefulness for the organisations involved is most tangible. However, for several reasons it is not obvious that M&E systems will gradually also ‘automatically’ integrate the higher levels of the ends-means chain and that the evaluation function in particular will be developed as strongly as the monitoring function. After all: the stimuli and positive preconditions that currently exist for M&E at implementation level (directly demonstrable usefulness; traditional focus on operational aspects; pressure from DGD; relatively minor set-up and
implementation requirements) do not or hardly exist for M&E with respect to outcomes and impacts. Hence, despite the progress of recent years in the development of M&E systems, increasing evaluability – in the first place as regards other criteria than efficiency – is not at all a future certainty if policy and context remain unchanged.

**Results of the comparative analysis.** Evaluability scores *per country* do not differ much, which implies that other parameters are probably more important. Only *context* shows considerable differences, for obvious reasons, but the influence of context on technical evaluability is not of such importance as to constitute a serious impediment. Certain institutional context can, by the way, have both a positive and a negative effect on the various aspects of evaluability, which is shown from the Rwanda experiences. *Interventions with a ‘complex’ theory of change (TOC) and those with a ‘less complex’ one* do not show substantial differences in terms of evaluability. Interventions with a complex TOC even score a bit higher, possibly because the stakeholders in these interventions invest more in analysis and implementation of M&E systems and practice, and because they – be it justified or not – assume that the interventions involved are harder to finance and that it is more difficult to demonstrate their results. The evaluability scores *per funding channel* (type Belgian stakeholder) show larger differences than those for the nature of the interventions and the country. Differences are particularly significant between ‘bilateral/NGOs/trade unions’ on the one hand and ‘other stakeholders’ on the other hand, with better scores for the former, even though the latter group also shows examples of good practices. The predominant explanation for this finding could be lower external (from DGD) requirements for the ‘other stakeholders’ as regards intervention plan and implementation practice, which are enhanced by the fact that for part of the stakeholders in this group, development cooperation is not a primary task.

**Main Recommendations**

The individual stakeholders in the Belgian development cooperation each have an interest and a responsibility in aiming for better evaluability, and had best address the issue from a *jointly* defined and supported framework and guidelines, which each group then can work with from their own role and specific characteristics. The study team realises that some of the formulated recommendations are quite demanding, at least at first. As such they can only be applied properly if the management burden for the stakeholders in question can be reduced accordingly, preferably – in accordance with the vision of the Strategy Note on Development Results – by giving a different interpretation to the requirements as regards the specifications of intervention plans, implementation reports, etc., shifting the focus from operational levels (means, activities, output) to development *results* (outcomes, impact).

The *strategic recommendations* are as follows:

1. Systematic integration, by all of the stakeholders, of evaluability and evaluability assessment into their operations, treating them as a means to better development cooperation rather than as leverage for bureaucratic control or direction (by donors and/or in organisations). The aim may not and cannot be to achieve maximum evaluability; increasing evaluability must be a continuous concern but it has to have an adequate place in a specific context; there will always be a turning point at which the costs of achieving better evaluability are no longer in balance with the benefits.

2. The introduction of a coherent evaluability assessment as an important tool in the ex ante analysis for each evaluation, for the purpose of analysing and demonstrating the potential benefits of the evaluation and thus reaching a well-founded decision on whether or not an evaluation will be carried out.

3. To improve the preparatory phase of interventions, with a focus on quality rather than on ‘more of the same’. Given that this involves a very demanding process, it is important to aim for *gradual* improvement, supported in different ways: via a
customised framework (with incentives) from DGD, via high-quality studies and evaluations that can support the formulation (e.g. at the end of previous stages) and via a reduction of any administrative requirements and regulations (linked to intervention propositions and reports) that could hinder the effectiveness of the development.

4. Renewed attention for and upgrading of the outcome and impact levels across the entire intervention cycle (intervention design, M&E, etc.) via, among other things, a clear definition of these basic concepts and how they are implemented in intervention propositions and reports, and proper specification of TOC, indicating direct (intermediate) outcomes to which, based on a clear TOC, interventions may contribute on the one hand, and long term effects at society level on the other.

5. Further development of – often already well developed – M&E systems and practices, with the aim to achieve good evaluability of effectiveness, impact and sustainability. Such development is best done gradually, with means, instruments, capacity and experience increasing at simultaneous speed so as to be able to include and integrate more complex functions in due time.

6. In connection with the previous point it is important to create a framework in which these – ambitious – changes are not only facilitated, but also stimulated and appreciated. DGD has a crucial role in this context and in consultation with the other key stakeholders could (a) further revise and ease the current rules and regulations, instruments and procedures so as to steer them more towards the intended development effects; (b) create incentives to further develop the M&E function, particularly the evaluation function, thus better enabling stakeholders to carry out high-quality evaluations which also analyse sustainability and impact; (c) establish a fund for the financing of studies and evaluations at effect and impact level, whereby the initiative should preferably lie with the aggregate Belgian stakeholders in the development environment. The fund should finance joint exercises involving several interventions from different stakeholders and implement studies and evaluations which either exceed the means and capacities of individual stakeholders, and/or are beyond their scope of interest.

7. The planned certification of M&E systems had best not be severed from a broader and more integrated approach, e.g. as planned in the screening of the NGAs. Rather than providing formal and standardised certification, the best line of action – in keeping with the past approach – would be to develop good incentives for increasing the quality and management of interventions and M&E systems in particular, which incentives would be based preferably on processes already in motion among the different stakeholders, such as the development and application (by the stakeholders, if necessary with external guidance) of a diagnostics instrument which enables the stakeholders (and DGD) to acquire better understanding of the strengths and weaknesses of their M&S system, to develop a tailor-made results based action plan for improving their M&E system. To (permanently) incorporate these important functions, DGD must continually have the necessary human and other resources.

The study concludes with a few operational recommendations which in part involve a concrete specification of the strategic recommendations:

8. Intervention designs must give more attention to a better description (differentiation) of target groups and the development of a good baseline; M&E policies and their translation into practice must be improved, with more attention for the development of an evaluation function and complementing one's own M&E policy with that of local stakeholders.

10. The M&E function at intervention level must be developed further, focusing particularly on the integration of the individual M&E components into a coherent system and more involvement – based on the principle of subsidiarity – in M&E among the stakeholders;

11. A more conscious approach is necessary for optimising the final use of evaluations, among other things by enhancing the involvement, at all stages, of the end users in evaluations and by planning and carrying out evaluations from a
portfolio approach, which, at an evaluation level, implies a differentiated approach of the different interventions.
4.3 Comparison of evaluability at stakeholder level ........................................... 58

5 Conclusions and recommendations ............................................................. 61
  5.1 Main conclusions .................................................................................. 61
  5.1.1 Synthesis of the main results and findings ........................................... 61
  5.1.2 Analysis ......................................................................................... 66
  5.2 Recommendations ............................................................................. 68
    5.2.1 Strategic recommendations ........................................................... 68
    5.2.2 Operational recommendations ....................................................... 72

Annexes ........................................................................................................... 75
List of tables

Table 1: Usefulness and importance of evaluability assessment ........................................ 17
Table 2: Summary presentation of the study framework ..................................................... 20
Table 3: Evaluability index per DAC criterion and component for the 40 interventions ... 24
Table 4: Principal results as regards the underlying analysis ........................................... 27
Table 5: Main results for intervention logic and theory of change .................................. 30
Table 6: Main results with respect to the proposed M&E system .................................... 32
Table 7: Main results for consistency and adaptation of the intervention logic and theory of change .............................................................................................................. 36
Table 8: Main results for the availability of basic information about the intervention implementation .................................................................................................................. 37
Table 9: The M&E system in practice .............................................................................. 41
Table 10: Findings with respect to the attitude of key stakeholders ................................. 46
Table 11: Overview of the findings on the broader context ............................................. 49
Table 12: Overview of the evaluability scores per country .............................................. 54
Table 13: Overview of the evaluability scores for interventions with a complex and less complex TOC ................................................................................................................ 56
Table 14: Overview of the evaluability scores per stakeholder ...................................... 58
## List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>APEFE</td>
<td>Association for the Promotion of Education and Training Abroad</td>
</tr>
<tr>
<td>BTC</td>
<td>Belgian Technical Cooperation</td>
</tr>
<tr>
<td>CUD</td>
<td>Cooperation from Universities to Development</td>
</tr>
<tr>
<td>DAC</td>
<td>Development Assistance Committee</td>
</tr>
<tr>
<td>DRC</td>
<td>Democratic Republic of the Congo</td>
</tr>
<tr>
<td>DGD</td>
<td>Directorate General for Development Cooperation and Humanitarian Aid</td>
</tr>
<tr>
<td>DFID</td>
<td>Department for International Development</td>
</tr>
<tr>
<td>FGD</td>
<td>Focus Group Discussion</td>
</tr>
<tr>
<td>GC</td>
<td>Guidance Committee</td>
</tr>
<tr>
<td>G&amp;D</td>
<td>Gender and Development</td>
</tr>
<tr>
<td>IMEP</td>
<td>Independent Monitoring and Evaluation Project</td>
</tr>
<tr>
<td>IOB</td>
<td>Institute for Development Policy and Management</td>
</tr>
<tr>
<td>ITG</td>
<td>Instituut voor Tropische Geneeskunde (Institute for Tropical Medicine)</td>
</tr>
<tr>
<td>M&amp;E</td>
<td>Monitoring and evaluation</td>
</tr>
<tr>
<td>MIS</td>
<td>Management Information System</td>
</tr>
<tr>
<td>NGA</td>
<td>Non-governmental actor(s)</td>
</tr>
<tr>
<td>NGO</td>
<td>Non-governmental organisation(s)</td>
</tr>
<tr>
<td>OECD</td>
<td>Organisation for Economic Cooperation and Development</td>
</tr>
<tr>
<td>PCM</td>
<td>Project Cycle Management</td>
</tr>
<tr>
<td>SEO</td>
<td>Special Evaluation Office</td>
</tr>
<tr>
<td>ToC</td>
<td>Theory of Change</td>
</tr>
<tr>
<td>TFD</td>
<td>Technical and Financial Dossier</td>
</tr>
<tr>
<td>UN</td>
<td>United Nations</td>
</tr>
<tr>
<td>VLIR-UOS</td>
<td>Flemish Interuniversity Council, University Development Cooperation</td>
</tr>
<tr>
<td>VVOB</td>
<td>Flemish Association for Development Cooperation and Technical Assistance</td>
</tr>
</tbody>
</table>
1. Introduction

1.1 Background of the study

The initiative of the SEO (Special Evaluation Office) to carry out a study about evaluability was based on several motives\(^1\). Evaluations have become increasingly important in the management cycle in the last decade, both at the intervention level and higher\(^2\). The role of evaluations has become clearer, with the focus increasingly on the different objectives they may target, such as accounting for the use of received funds; learning from experiences from the past to improve in the future; and support for future policy formation and practices.

Given the increasing importance of evaluations in development cooperation – evaluations are deemed of crucial importance, are mandatory in many cases and, in combination with monitoring have often grown into complete systems – it is important to pay attention to the relevance and quality of evaluations and the follow-up of their results. A first and important, but often forgotten step is to check whether all the conditions have been met to carry out a high-performance evaluation; in other words, whether the evaluation meets an existing need, whether it is feasible in the given context and with the available funds, and whether the use of those funds is justified in light of the benefits the evaluation is expected to generate. Furthermore it is important to keep in mind that both the meaning of the word and the practice of evaluation have eroded: despite the gain in evaluation experience, in many of the ‘evaluation’ initiatives one may question whether the minimum requirements that actually make an evaluation, have been fulfilled\(^3\). The erosion is based on a lack of expertise, but also has to do with the fact that evaluations have started forming part of the ‘system’ and as a result, have been initiated and carried out as (too much of) a routine.

Besides these international developments, the experiences in Belgium are another reason to dedicate a study to evaluability. After all: experiences with external evaluations, but also with monitoring, suggest that in many cases, the initial conditions for proper evaluation are absent or only partly present. As such, one may assume that these evaluation exercises can fulfil their ambitions only partially. On the other hand, many Belgian development organisations have been investing in the development of their M&E systems in recent years, which – in principle – should have led to better evaluability. However, there is not much data available on the actual usefulness of these systems and particularly the results of these evaluations. Evaluability is an interesting concept in this context, as it can support the decision-making regarding the desirability of evaluation (of a project, a programme, etc.).

From the above findings, the concept of evaluability recently has become more popular, where it didn’t get much attention in the past. SEO therefore decided to dedicate a study to evaluability and carried out the first two stages itself. Those two stages involved the development of an analytical framework of theoretical evaluability, and a study of the

\(^1\) Chapter 2 explains the concept evaluability in further detail.

\(^2\) In this study, the term intervention is used for the aggregate of all the projects, programmes, instruments, etc. of the Belgian development cooperation.

\(^3\) From a methodological viewpoint, evaluations are expected to deliver reliable, valid and useful results. See also chapter 2.
theoretical evaluability of 43 interventions. The report of these two stages (of stage 2 in draft) was made available to us at the start of our study

1.2 Basic characteristics of this study and its context

First and foremost it is important to realise that the nature of this assignment differs substantially from other contracts assigned by the SEO.

- This assignment is a study, not an evaluation; therefore, the aim is not to meet the classic goals of an evaluation, but to meet a few clearly defined objectives (see 1.3 below) of the study which were formulated with a view to the future.
- As mentioned earlier, the SEO already carried out the first two stages of the study, namely the development of an analytical framework of the theoretical evaluability and a study of the theoretical evaluability of 43 interventions.
- As mentioned in the Specifications, in this study the SEO by way of exception is also a stakeholder; while in ‘traditional’ assignments issued by it, the SEO is involved mainly as the Senior Official, it has a greater interest in this study since the realisation of the objectives is of direct significance to it.
- The existing literature on the subject of evaluability shows that the concept can be – and is – used in different ways. For this reason it is important to provide a clear definition of the concept of evaluability. The definitions in the specifications (and in literature) refer to evaluability in the narrow sense of the word as well as to evaluability assessment, while not always making a strict distinction between the two (i.e. the definition and the instrument). Although the definitions mentioned have helped in identifying the boundaries of the study, they required further explanation in the course of the study (see chapter 2). In this context it was not the intention to have an exhaustive discussion on the concept, but rather to reach pragmatic agreement on the working definition to be used throughout the study.

1.3 Objectives and scope of the study

Normally in a study or evaluation, the objectives and research area are defined ex ante and they remain – almost – unchanged while the study is being carried out. That was not the case in this study: the original objectives and research area as formulated in the study specifications were extended later on.

The objectives as formulated in the specifications of the assignment (ref. S4/2014/01) can be summarised as follows⁴:

- The study's objective is threefold:
  - to determine, together with the partners involved, the required, adequate and realistic conditions that facilitate objective evaluation of cooperation interventions;
  - to learn lessons that are useful for harmonisation and certification of the evaluation systems of the stakeholders provided by the legislator; and
  - to check to which extent the above-mentioned evaluability conditions coincide (or fail to coincide) with the conditions necessary for monitoring and results-oriented management of the interventions and with the legal and regulatory framework mentioned.
- The general objective is to – contribute to – making all future actions evaluable in the not too far future.

⁴ See point B3, p. 25 of the referred document, which can be found in Appendix 1. It is important to note that particularly the second and the third specific objective were based largely on initiatives that were planned at the launch of the study. For several reasons, the realisation of these initiatives did not work out as planned.
Furthermore it was decided that the results of the study in several ways had to be useful for SEO, the departments of the DG-D and the Attachés, as well as the partner organisations of the bilateral and non-governmental cooperation.

In the start-up phase of the study there was not enough clarity about the above-mentioned objectives and, inherently, distrust among some stakeholders as to the actual intentions and intended and unintended effects of the study; much of this was ventilated during the first meeting of the Steering Committee.

For this reason, in the first stage the study team has given ample attention to creating support for the study, among other things through better specification of the concept of evaluability. As a result thereof, the main objective of the study – to contribute to better evaluability of the actions of the Belgian cooperation – was slightly modified and the attention focused on the potential of analysing evaluability and improving the management of interventions, as well as determining the usefulness of an evaluation in a specific situation and in general. In addition, the above-mentioned specific objectives were given ample attention insofar as possible and relevant.

Finally, the study team’s suggestion to extend the initial focus – defined in the specifications of the study – on practical evaluability to a focus on both theoretical and practical evaluability, was accepted5.

1.4 Structure of this synthesis report

The present synthesis report has the following structure: This introductory chapter is followed by a chapter which pays attention to the concepts of evaluability and evaluability assessment. This chapter also seeks to demonstrate the importance and practical usefulness of these concepts, and answer the question why it can be interesting to carry out an evaluability test. It also briefly explains the research methodology and study framework used. Chapters 3 and 4 discuss the main findings of the study, while Chapter 5 contains the most important conclusions and recommendations. A number of appendices complete the report: the terms of reference; a description of the methodology; the study framework; the list of 40 analysed interventions; the bibliography; the list of contacted persons; and two technical appendices, namely an explanation of the statistical analyses and an extensive score table.

---

5 These concepts are explained further in Chapter 2.
2. Basic concepts, study framework and approach

2.1 Evaluability and evaluability assessment

2.1.1 Definitions

In the literature, the OECD/DAC definition of evaluability (2002: p.21) is the one most frequently used: "Evaluability is the extent to which an activity or a programme can be evaluated in a reliable and credible fashion". Furthermore, evaluability and evaluability assessment are often closely linked. The latter is then described as 'an instrument which allows to determine whether in a certain situation evaluation should be carried out'. The link implies that it is important to use evaluability not only as a concept, but also to link it to evaluation practices and the necessity to check whether evaluation is justified and feasible and capable of providing useful information. This means that the contribution of an evaluation to the improvement of the management and the results of the actions are included in the assessment of the evaluability.

Before continuing it is important to briefly analyse the OECD/DAC definition, since it is deceptively simple and important implications could be easily missed as a result. The concepts reliable and credible are focal points that are frequently used in the context of evaluations, but not always specified (as it isn't in the OECD/DAC definition):

- An evaluation is reliable when its study results are stable and consistent. This means that if the evaluation study would be done again, it would produce the same findings. Elements that affect reliability include the quality of the research methods and implementation, the scope of the research, the independence of the information and the researchers.
- An evaluation is credible when the research results are considered valid and relevant by the most important parties (stakeholders). Elements that may affect the credibility include: attention and sensitivity for context-specific (economic, cultural, social) dimensions of the intervention; broad data collection and triangulation between sources and an extensive description of the data collection process; transparency and independence of the evaluation process; reliability of the measuring instruments; consistence and coherence of the findings and between the findings and conclusions.

An important distinction is often made between theoretical evaluability and practical evaluability. Theoretical evaluability refers to evaluability in principle as it can be derived from the design of the intervention, without taking practice into the equation. Practical evaluability on the other hand will check, for instance, whether the data concerning the progress of an intervention is in fact collected and whether M&E systems are actually

---

6 "Evaluability is the extent to which an activity or project can be evaluated in a reliable and credible fashion"; see OECD/DAC (2002) Glossary of Key Terms in Evaluation and Results Based Management. Paris: OECD/DAC, p. 21.
being used, so that they can contribute to improving management and results. Thus, the link to – practical – usability of the evaluation is quickly established.

This study approaches evaluability mainly as a **continuum**. No intervention is perfectly evaluable in all its aspects, while on the other hand it is quite unlikely that there are interventions that cannot be evaluated at all. Furthermore, an evaluation may entail several aspects of an intervention: the five classic DAC criteria: relevance, efficiency, effectiveness, impact and sustainability, but other criteria as well. However, it is not necessary that all evaluations pay attention to all these criteria at the same time. In addition, evaluability can have an association with the objective of the evaluation. While evaluations may have several objectives, such as accountability, learning, and supporting the future policy, it is not necessary to include them prominently at the same time in all evaluations. For instance, if learning is the main objective, the requirements in the field of evaluability will not be the same as when accountability is the focal point. And finally, evaluability also depends on the evaluation policy of the organisation. It may opt to use the available means for strategic evaluation of innovative interventions, for instance, or for evaluation of all interventions; it may also opt for focussing on independent evaluations, concentrating on accountability, or on internal evaluations with a strong learning character.

### 2.1.2 Why is assessment of evaluability important?

First of all, it is important to state that determining the evaluability of an intervention does **not** serve to determine the (development) value of that intervention. It is perfectly possible that highly valuable interventions are hard to evaluate. In other words: there is no connection between the extent of evaluability and the (development) value of an intervention. Evaluability assessment has more to do with the nature of the intervention and the context in which it is being carried out. For instance: the challenges to evaluate interventions aiming at empowering women generally will be more substantial than to evaluate interventions about potable water supply. This can be explained from the fact that **empowerment** can have totally different meanings for the stakeholders; that it involves an iterative process with various dimensions and stages, etc., whereas it is relatively simple to get agreement on the meaning of **potable water supply**. Context also plays an important role: it is difficult to carry out evaluations in conflict areas or in interventions where the parties involved are at daggers drawn.

So what is the importance and usefulness of an assessment of evaluability? Although the extent of evaluability and the feasibility of an evaluability assessment depend on different factors (see above), the question does allow for a number of general answers:

- **Studies about the quality of evaluations show that in many cases, it is below par or that evaluations have little added value compared to their costs; furthermore, it appears that the effective use of evaluation results leaves room for improvement. A good assessment of evaluability may offer a partial solution to these problems.**

- **If the intention is to carry out an evaluation, the assessment of the evaluability can usually be carried out at a fraction of the total evaluation budget; thus, one can avoid that funds are wasted, for instance on evaluations that are not required or feasible, or that are designed badly. Such analysis would seem required especially in the implementation of complex evaluations, whereas it would be less necessary for well-planned and well-performed interventions with well-functioning M&E systems.**

- Furthermore, the specific importance and usefulness of an evaluability assessment will depend on when they are carried out in the intervention cycle (see also table 1 below):

  - **in the preparation phase of an intervention** an evaluability assessment will focus mainly on the intervention design: is it coherent and complete? Have all assumptions (also the implicit ones) been included? Does it reflect the...
underlying analysis? Etc. Such evaluability assessment sees at the theoretical evaluability defined above. A good or improved intervention design obviously will have a positive effect on the quality of the intervention (including the M&E system) and later evaluations;

- at the start of an intervention, or just before, an evaluability assessment based on the intervention design and the general intervention proposition will be able to supply valuable elements to the development of an M&E system;

- during the implementation of an intervention an evaluability assessment may provide indications of the desirability, timing and feasibility of an evaluation by checking to which extent the conditions for adequately performing an evaluation have been fulfilled; e.g.: what is the attitude of the parties involved? Such assessment may also verify whether the intervention design is adequately in tune with external developments or the acquired insights and propose suggestions if necessary;

- at the completion phase of an intervention the assessment focus will be quite similar and provide indications as regards the desirability and feasibility of an evaluation, the main difference being that at this stage improvements are no longer possible; the main focus will lie on the challenges with respect to adequate performance of the evaluation and how to deal with them, and on any alternatives available in this respect;

- after termination of an intervention the focus of the assessment will be in line with that at the completion phase of an intervention, with the feasibility of contacting the stakeholders in the intervention as an important additional aspect.

Furthermore, it is clear that a systematic examination by a development organisation of the evaluability of interventions will contribute to better intervention practices and to development and implementation of a proper evaluation policy and to careful consideration of all the interventions to be evaluated and of the different evaluation objectives. It also makes it possible to make a clear distinction between evaluability assessment and desirability to evaluate
Table 1: Usefulness and importance of evaluability assessment

<table>
<thead>
<tr>
<th>Intervention cycle phase</th>
<th>Purpose of the assessment</th>
<th>Focus of the assessment</th>
<th>Intended result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formulation of the intervention</td>
<td>Improvement of the intervention design</td>
<td>Quality of the intervention design</td>
<td>Improvements and additions to the intervention design</td>
</tr>
<tr>
<td>Start of the intervention</td>
<td>Supplying input for developing an M&amp;E system</td>
<td>Availability and quality of the information and data collection</td>
<td>Improvement of the M&amp;E system (contents and process)</td>
</tr>
<tr>
<td>Implementation and completion</td>
<td>• Decide whether evaluation takes place now or later • Check that intervention is in line with developments</td>
<td>• Availability, being up-to-date and quality of data • Position/opinion of parties involved • Situation in broader environment</td>
<td>• Insight into degree of difficulty of implementing the evaluation at this point • Proposing alternatives as regards timing and contents • Modification of the intervention design and M&amp;E system</td>
</tr>
<tr>
<td></td>
<td>Supplying input for the design in a planned evaluation</td>
<td></td>
<td>• Insight in the options as regards evaluation objectives, main questions, approach and required expertise</td>
</tr>
<tr>
<td>Post-intervention</td>
<td>Check whether evaluation in a reliable and credible manner is possible</td>
<td>• Availability risk and quality of the data about intervention effects • Feasibility of contacting the parties involved</td>
<td>• Insight into degree of difficulty of implementing the evaluation • Insight into options to evaluate (aspects of) impact and sustainability • Insight into the desirability of the evaluation</td>
</tr>
</tbody>
</table>

2.2 The actual study framework

Based on the objectives of the study we have tried to design a realistic study framework, taking account of the available funds and the requirement of examining some 40 interventions. In other words: proposing an excessively detailed framework was useless if it would appear inapplicable later on. In addition we wanted to avoid a ticking the boxes approach that could follow from a too detailed framework and which would undoubtedly have a negative effect on the quality and relevance of the exercise. These considerations have led to a scenario which allowed four days per intervention (including field work) for determining evaluability, which is in keeping with the maximum number of days that can be made available for the analysis of the evaluability of an intervention.

On the other hand, we could not but admit that evaluability comprises multiple dimensions and aspects, such as theoretical and practical evaluability. In our opinion we

---


9 In this context we can state that in the context of the Belgian cooperation, evaluability assessments actually hardly take place. However, very often the preparatory path of evaluations contains elements that are also included in evaluability assessments.
reached a good compromise between reality and completeness by applying a hierarchy in the individual parts of the study. A couple of central concepts/criteria, to which we will be referring as components, were specified into clusters of underlying factors or items, as we call them in this study; this approach is comparable to the one used in evaluations whereby main questions are operationalised into assessment criteria, which in turn are specified into indicators.

Furthermore, the study framework allows for scoring all of the components, which enables aggregation and comparison of/between interventions, components, stakeholders, countries, etc. The aggregated score can be considered an indicator of evaluability which will sit somewhere in the continuum between no evaluability and total evaluability. The development of the study framework was inspired mainly on the study performed by Rick Davies for DFID\(^\text{10}\). However, compared to Davies our study framework contains more detail. The framework consists of three parts, hereinafter referred to as 'dimensions':

- the analysis of the intervention design (including the underlying theory of change) which mainly (but not only) describes the theoretical evaluability. The analysis of the intervention design in a strict sense was used to also assess the quality of the planned M&E system;
- the analysis of the practice with respect to the implementation and management of the intervention and the context (including the use of human and financial resources);
- the analysis of contextual factors; these factors can play a role both in the field of the intervention design and in the field of the data-generating systems, but also in the (future) implementation of the evaluation itself.

The first part mainly involves the theoretical evaluability (but not solely: the planned M&E system does not form part of it); the second part examines the practical evaluability but does not comprise it in full; the third part contains elements that may influence both the theoretical and the practical evaluability\(^\text{11}\).

Table 2 gives an overview of the study framework and of the items scored for each of the five DAC evaluation criteria. It is important to note that the DAC definition of the evaluation criteria was used as the basis for working out the framework (see framework).

The study framework comprises 3 dimensions in total, which are subdivided into 9 components and 62 items, all of which were scored\(^\text{12}\). Scoring was based on a five-point scale for each component and each item; for the dimensions it was based on a weighted average calculated for the components\(^\text{13}\). The complete study framework is included in appendix 3.

---


\(^{11}\) As indicated above, the SEO carried out a study on the theoretical evaluability, with the aim to serve as input for the present study and focussing on practical evaluability. However, since the sample of the SEO study phase was not exactly the same as that of the present study, it became clear that we would have to include theoretical evaluability in our study as well. That was the only option for closely examining the different aspects of practical evaluability as well as the links with theoretical evaluability.

\(^{12}\) At a later stage in the process, one of those nine components was partially struck out and partially incorporated into another component. The table reflects the modified study framework.

\(^{13}\) See Appendix 2 for additional explanation.
Framework 1: The DAC evaluation criteria and their interpretation criteria for this study

Relevance: the extent to which the objectives of a development intervention are consistent with the requirements of the beneficiaries, the needs of the country, the general priorities and the policy of the partners and donors. The study maintained this definition but in practice specifically examined the first two elements.

Effectiveness: the extent to which the objective of an intervention has been achieved (or is expected to be achieved), taking into account its relative importance. The study copied this definition, but has assessed effectiveness mainly at the level of the specific objective (effect, outcome) of the interventions.

Efficiency: the extent to which the means and input (funds, expertise, time, etc.) of an intervention are converted into results. This study has interpreted efficiency in a likewise manner by looking at the way in which input is converted into activities and then leads to outputs.

Impact: the positive and negative, most important and secondary long-term effects produced by an intervention, directly or indirectly, deliberately or by accident. This study used this definition, but has also used the following concept for impact: ‘directly attributable effects (causality) of an intervention’.

Sustainability: the extent to which the benefits of an intervention continue after completion of substantial support. The study copied this definition.

---


15 In other DAC documents, results can be outputs, outcomes as well as impacts. In the compendium, DAC replaces results by the narrower definition ‘outputs’. We chose for the narrower definition, because it allows to make a clear distinction between effectiveness and efficiency, which was important for this study. It is important to realise that this was not the only option and furthermore, that the broader definition is preferred by most experts in the area of efficiency, for instance Palenberg who carried out a study for BMZ which is often cited in the sector: Palenberg, M., Tools and Methods for Evaluating Efficiency in Development Interventions, BMZ Evaluation Division – Evaluation Working Papers, 2011, 131 p.

16 In this study, the concept impact is used in two different ways:
- as an evaluation criterion in the strictest sense, in which case it is distinguished from, among other things, ‘effectiveness’ and refers to results ‘above’ the outcome level; as such, ‘impact’ refers to the general objective(s) as formulated in the logical framework of an intervention;
- in the scope of a discussion of (the practice of) ‘impact evaluations’ as a specific evaluation method, in which case the focus is on effects that are ‘directly attributable’ to an intervention (causality). These effects can be situated at the outcome level as well as the higher levels in the chain. In that case an impact evaluation seeks to distinguish the effects attributable to an intervention (causality) from the effects caused by other factors (external factors or the set-up and instruments of the evaluation itself). This interpretation is in line with the one used in the SEO impact evaluation: Impact meten: zoektocht naar de Graal, Ex-post impactevaluatie van vier projecten van de gouvernementele samenwerking (Measuring impact: search of the Grail, ex post impact evaluation of four projects of the governmental cooperation). This study defines impact as: the effects arising from the project at a general level. Incidentally, the impact evaluations carried out via SEO involve mainly outcome evaluations.

Obviously this dual use of the concept ‘impact’ could cause confusion; we have tried to minimise it in the study.
Table 2: Summary presentation of the study framework

<table>
<thead>
<tr>
<th>Dimensions/Components</th>
<th>Relevance</th>
<th>Effectiveness</th>
<th>Efficiency</th>
<th>Impact</th>
<th>Sustainability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Analysis of the intervention design</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1 Underlying analysis (7)</td>
<td>(7)</td>
<td>(4)</td>
<td>(0)</td>
<td>(0)</td>
<td>(0)</td>
</tr>
<tr>
<td>1.2 The intervention logic and theory of change (8)</td>
<td>(1)</td>
<td>(6)</td>
<td>(5)</td>
<td>(5)</td>
<td>(3)</td>
</tr>
<tr>
<td>1.3 The proposed M&amp;E system (9)</td>
<td>(5)</td>
<td>(7)</td>
<td>(9)</td>
<td>(7)</td>
<td>(7)</td>
</tr>
<tr>
<td>1.4 The consistency and adaptation of the intervention logic and theory of change (3)</td>
<td>(3)</td>
<td>(3)</td>
<td>(3)</td>
<td>(3)</td>
<td>(3)</td>
</tr>
<tr>
<td>2. Practice with respect to implementation and management of the intervention and the context</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1 Basic information concerning the intervention implementation (11)</td>
<td>(5)</td>
<td>(10)</td>
<td>(11)</td>
<td>(10)</td>
<td>(8)</td>
</tr>
<tr>
<td>2.2 The M&amp;E system in practice (12)</td>
<td>(11)</td>
<td>(12)</td>
<td>(12)</td>
<td>(12)</td>
<td>(12)</td>
</tr>
<tr>
<td>3. The evaluation context</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1 Attitude of the key stakeholders (9)</td>
<td>(9)</td>
<td>(9)</td>
<td>(9)</td>
<td>(9)</td>
<td>(9)</td>
</tr>
<tr>
<td>3.2 The broader context (3)</td>
<td>(3)</td>
<td>(3)</td>
<td>(3)</td>
<td>(3)</td>
<td>(3)</td>
</tr>
<tr>
<td>3.3 Practical elements (2) (*)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggregate score</td>
<td>(44)</td>
<td>(54)</td>
<td>(52)</td>
<td>(49)</td>
<td>(45)</td>
</tr>
<tr>
<td>4. Suggestions with respect to the M&amp;E system used and future evaluations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Feedback on our analyses and suggestions from the persons involved</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(*) This part is mentioned only as a reminder since the parameters and scores for it are inevitably linked to the organisation of a concrete evaluation.

2.3 Approach

2.3.1 Sample determination

Pursuant to the requirements laid down in the specifications, 40 interventions were analysed in 4 countries (Belgium, Benin, DRC and Rwanda); for each country, 10 interventions were selected. They all involved interventions that were in an advanced state of implementation, some via a previous phase. The sample determination was not at random, but aimed at over-representation of smaller sectors and stakeholders and of a-typical interventions (e.g. college grants). In view of a later analysis, the sample determination used two parameters in addition to ‘country’, namely ‘complex v. less complex interventions’ and ‘the intervention channel’ (direct bilateral cooperation, indirect cooperation via NGOs and trade unions, cooperation via other stakeholders).

---

17 Bijlage 2 bevat meer informatie m.b.t. de aanpak en methodologie van de studie.
18 A list of all 40 projects is included in appendix 4.
19 After a long process of internal reflection, this parameter was given a specific function: rather than introducing a distinction based on complex and less complex sectors, we chose for distinguishing between interventions with a complex TOC and interventions with a less complex TOC. If, for instance, a specific intervention in a ‘less complex’ sector, was nonetheless assessed as ‘complex’ based on the underlying TOC, it would end up in the latter category. It is equally important to note that the distinction that was applied also differs from that between complex and less complex interventions as such, in which other factors also play a part, such as the size of the intervention, the number of stakeholders, etc.
The final sample consisted of:
- 10 interventions in each of the 4 countries
- 24 (60%) complex and 16 (40%) less complex interventions
- 10 (25%) bilateral interventions, 20 (50%) interventions carried out by NGOs or trade unions, and 10 (25%) interventions carried out via other stakeholders (APEFE/VVOB, universities, BIO, …)

### 2.3.2 Data collection and analysis

Data collection was carried out via a combination of secondary and primary data collection. In practice the following methods were most important:

- study and analysis of basic documents (intervention proposition; baseline (if carried out); implementation reports and evaluations; documents concerning M&E);
- interviews with the key stakeholders in the implementation of the interventions;
- discussions in focus groups (at a stakeholder level and intervention level).

Based on the collected data the analysis framework was filled out for each intervention by scoring the components and items. For the purpose of reducing partiality, each intervention (with the exception of 5 interventions in Belgium) was scored by at least two evaluators; where different scores were found, agreement was sought through discussion. The scores were then entered into tables, which were subjected to a number of simple and more complex statistical analyses (see appendices 2 and 7), which formed the basis of the findings presented in the next two chapters.

### 2.3.3 Study phases: summary of the main steps

From the start to the completion of the final report, the study took about eight months (from late February to mid October 2015). The start-up took a relatively large amount of time because of the rather a-typical subject and initial resistance from a number of stakeholders (see above). A relatively large amount of time was spent on discussing the basic concepts and the study approach, among other things via the development of an explanatory paper. The study of the basic documents took place from March to May and included a pilot field study in Rwanda in April-May and visits to Benin and DRC in May-June. The projects in Belgium were analysed mainly in June and July. Editing the draft synthesis report took place in July and August, after which the draft final report was discussed in the Steering Committee at the end of September.

Meetings with the Steering Committee were held at crucial moments during the process: at the start-up of the study, after the first field visit, and after finishing the draft synthesis report.
3. Main general findings

General Overview

This chapter looks into the findings from a general perspective, i.e. as regards the 40 interventions as a whole. This general analysis produces general conclusions, which later on (in chapter 4) will be further detailed per country involved, intervention channel or level of complexity (of the TOC) of the intervention.

The table below (table 3) provides an overview of the average scores and the evaluability index for the eight components and five evaluation criteria that were briefly introduced above (see 2.2). As indicated before, each component was operationalised into a number of items (the number between brackets); furthermore, the scores per component are not the average for the item scores per component, but for the scores attributed by the researchers at the level of each component and for each of the 40 interventions.

Each component (and each item) was given a score from 1 to 5, 1 representing a bad performance and 5 a good performance. In this context, the first component (the underlying analysis) could be given a score of 1 if such analysis was not carried out/not written, whereas it would get a score of 5 if the analysis was good and complete and supported the objectives of the intervention. A score of 3 would then mean that the analysis had been developed properly, but incomplete (e.g. no analysis of the gender relations). In some cases (the hatched cells) no score was attributed because there was no (or only a very indirect) relation between the component and the evaluation criterion. The more specific results at a component and item level will be presented and discussed in detail further down in this chapter.

In this table and several similar tables further down in the report, we have used colours to distinguish the scores. The colours provide a first and very general insight into the relative evaluability of the interventions. The higher the score or index, the better the evaluability. It is important to realise that the value of the index cannot be interpreted in an absolute sense, but can be used only in a relative sense, namely to compare different components and criteria among themselves and do the same with the 40 interventions. Thus it could be established that of all of the components, the ‘broader context’ has the highest score, and that ‘the proposed M&E system’ has the lowest score. We also established that ‘impact’ is the criterion with the lowest score (for practically every component). The table also allows us to quickly find the weakest score: the proposed M&E system at impact level.
### Table 3: Evaluability index per DAC criterion and component for the 40 interventions

<table>
<thead>
<tr>
<th>Item</th>
<th>Relevance</th>
<th>Effectiveness</th>
<th>Efficiency</th>
<th>Impact</th>
<th>Sustainability</th>
<th>Evaluability index</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dimension 1 (intervention plan)</strong> (*)</td>
<td>3.11</td>
<td>3.22</td>
<td>3.29</td>
<td>2.52</td>
<td>3.15</td>
<td>3.16</td>
</tr>
<tr>
<td>1.1 The underlying analysis (7)</td>
<td>3.65</td>
<td>3.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2 The intervention logic and theory of change (8)</td>
<td>3.00</td>
<td>3.50</td>
<td>2.10</td>
<td></td>
<td>3.25</td>
<td>2.96</td>
</tr>
<tr>
<td>1.3 The proposed M&amp;E system (9)</td>
<td>2.45</td>
<td>2.83</td>
<td>2.85</td>
<td>1.95</td>
<td>2.46</td>
<td>2.51</td>
</tr>
<tr>
<td>1.4 Consistency and adaptation of the intervention logic and theory of change</td>
<td>3.55</td>
<td>3.44</td>
<td>3.52</td>
<td>3.50</td>
<td>3.75</td>
<td>3.54</td>
</tr>
<tr>
<td><strong>Dimension 2 (implementation practice)</strong></td>
<td>2.99</td>
<td>3.25</td>
<td>3.51</td>
<td>2.33</td>
<td>2.73</td>
<td>2.96</td>
</tr>
<tr>
<td>2.1 Basic information as regards the intervention implementation (11)</td>
<td>2.95</td>
<td>3.20</td>
<td>3.53</td>
<td>2.13</td>
<td>2.62</td>
<td>2.88</td>
</tr>
<tr>
<td>2.2 M&amp;E system in practice (12)</td>
<td>3.03</td>
<td>3.30</td>
<td>3.50</td>
<td>2.53</td>
<td>2.85</td>
<td>3.04</td>
</tr>
<tr>
<td><strong>Dimension 3 (context)</strong></td>
<td>3.89</td>
<td>3.98</td>
<td>3.98</td>
<td>3.84</td>
<td>3.89</td>
<td>3.91</td>
</tr>
<tr>
<td>3.1 Attitude of the key stakeholders (9)</td>
<td>3.68</td>
<td>3.70</td>
<td>3.70</td>
<td>3.58</td>
<td>3.63</td>
<td>3.66</td>
</tr>
<tr>
<td>3.2 Broader context (3)</td>
<td>4.10</td>
<td>4.25</td>
<td>4.25</td>
<td>4.10</td>
<td>4.15</td>
<td>4.17</td>
</tr>
<tr>
<td><strong>General score for evaluability (°°)</strong></td>
<td>3.26</td>
<td>3.38</td>
<td>3.52</td>
<td>2.70</td>
<td>3.13</td>
<td>3.23</td>
</tr>
</tbody>
</table>

(*) The scores for evaluability at dimension level are unweighted averages of the scores at component level.

(°°) For the general evaluability score a weighted average was calculated in which dimensions 1 and 2 are weighed equally and dimension 3 counts for half of each of the other dimensions.

The table above allows us to formulate a number of **general findings** with respect to the different dimensions and components of the study framework and the evaluation criteria.

- The first thing we notice is that with the exception of dimension 3 (context), no dimension, component or criterion has a high score: the maximum scores are around 3.50 and many scores are in the yellow zone, i.e. below the average level on our scale. This indicates that in general, there is substantial room for improvement.
- As regards the three **dimensions**, we can see that dimension 3 (the context) scores remarkably better than the two other dimensions and that out of the other two, the implementation practice (dimension 2) has a weaker score than the intervention plan (dimension 1)\(^\text{21}\).
- In the area of the **components** we again see a high score for the two components that form part of the context; the weakest components are the proposed M&E system (dimension 1) and the basic information regarding intervention implementation (dimension 2).
- With respect to the **evaluation criteria**, the general evaluability score is lowest for the impact criterion; furthermore, sustainability also has a relatively weak score, whereas effectiveness and particularly efficiency show relatively high scores. In fact the practically equal scores – for all of the evaluation criteria – of

---

\(^{20}\) Not all of the 40 interventions could be scored for each DAC criterion. Table A1.2 in Appendix 7 shows when and to what extent a smaller number of interventions could be scored.

\(^{21}\) At a later stage (part 3 of this chapter) we will, however, add important notes to the good score for the components and items falling under ‘context’.

---

Evaluability Study – final report
the third dimension (context) lower the relative differences between the five criteria, and particularly between impact on the one hand and the four other criteria on the other. In other words: if dimension 3 (the context) were taken out of the equation, the differences between the evaluation criteria would be more outspoken. Other remarkable findings:

- The evaluability of efficiency scores relatively well, with a reasonably good (green) score for almost all of the components. The only weaker component here is the proposed M&E system (in the intervention plan), but this is compensated by a rather good score for the M&E system in practice. The relatively good score for efficiency illustrates what we will describe hereafter as a relatively good M&E practice at an operational level.
- As regards the evaluability of effectiveness, the majority of the scores are in the ‘reasonable’ (light green) zone, except for 1.2 and, linked to it, 1.3 (intervention logic and TOC; proposed M&E system). The relative weakness indicates too little attention for, particularly, the outcome level of the intervention plan.
- As far as sustainability is concerned, the weak M&E system scores (1.3 and 2.2) indicate a lack of attention for sustainability in the M&E system.
- As far as the impact criterion is concerned, the table shows that the evaluability of impact has a very moderate to weak score for all of the components that can be linked to intervention management (dimensions 1 and 2)\(^2\). For this weak score, we will identify a number of interlinked reasons which in essence involve the limited attention given by all of the parties to the impact level. The low evaluability score for impact obviously is in line of what was to be expected given the methodological challenges of impact evaluation, but also, and more important in the scope of this study, the relatively limited attention given to the impact level by the stakeholders.

Finally, we have also performed a general check as to the distribution of the general evaluability scores of the 40 analysed interventions (see figure below). The figure indicates a quite normal distribution (from a statistical point of view), with the highest frequency of interventions sitting around the average score (3.23) of the evaluability index. Also remarkable is the relatively large spread (minimum score of 1.94 and maximum score of 4.33) and the relatively large number of interventions (20% of the sample) with a high score (4.00 or more).

\(^2\) In this context we disregard component 1.4, where the number of analysed interventions (10 in total) is too small to draw any conclusions.
3.1 Analysis of the intervention design

This part successively looks at: the underlying analysis (7 items); the intervention logic and theory of change (8 items); the proposed M&E system (9 items) and the consistency and adaptation of the intervention logic and theory of change (3 items).

3.1.1 The underlying analysis

The quality of the underlying analysis affects the quality of evaluations focussing on relevance and effectiveness; without a good analysis, including an analysis of the (situation of the) target groups, an evaluation will have a hard time demonstrating whether or not a specific intervention is relevant for the target groups and has the intended effect on those groups. A (possible) relation between the underlying analysis and the evaluability of other DAC criteria (efficiency, impact and sustainability), however is less likely. For this reason, this component – other than the other components – received scores for relevance and effectiveness only.

Seven items were discussed with respect to this component, which are included in table 4 below together with their scores.
<table>
<thead>
<tr>
<th>Table 4: Principal results as regards the underlying analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1. The underlying analysis</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>1.1.1 The target groups are clearly defined and described</td>
</tr>
<tr>
<td>1.1.2 The rationale of the intervention and the (problem) situation of the target group are clearly described</td>
</tr>
<tr>
<td>1.1.3 The role of the target group(s) is clearly described</td>
</tr>
<tr>
<td>1.1.4 The role of the major stakeholders (excluding target group) is clear</td>
</tr>
<tr>
<td>1.1.5 Gender analysis forms an integral part of the underlying analysis</td>
</tr>
<tr>
<td>1.1.6 The relation between the underlying analysis and the intervention objectives is clearly described</td>
</tr>
<tr>
<td>1.1.7 The relation between the rationale of the intervention and the sector policy of the partner country is clearly described</td>
</tr>
</tbody>
</table>

**Important remark:** a further detailed version of the table below (and of the following similar tables) can be found in Appendix 8, which contains a fundamental explanation for each score. This fundamental explanation accounts for the score and is based on an assessment protocol used by the researchers which had to ensure consistent and more objective scoring by the researchers. Readers interested in the fundamental content of each score are therefore advised to read Appendix 8.

Compared to the seven other components, the underlying analysis scores relatively well in general; the other components of part 1 and part 2 produce lower scores. Only the two components under the evaluation context (part 3) show higher scores (See Table 3). In many cases, the approach is the result of practices developed and formed over the years, and which entail a division of tasks between the Belgian organisation, its local representation (if any) and local stakeholders (see box 1).

**Box 1 : Underlying analysis and development of an M&E system in Benin Islands of Peace**

VECO has taken the strategic decision to specialise in supporting a few specific agricultural chains. The organisation has developed a number of stakeholder-specific procedures and instruments to guarantee coherence of its field actions. These instruments ensure that the contextual analyses are performed coherently in each zone and contribute to a process of capitalization.

With respect to the underlying analysis, VECO has created several documents (both manuals and analyses) which enable a thorough analysis of the context and chosen strategies (made concrete for Benin via a regional planning agency). Subsequently, agencies at country level negotiate the indicators with the stakeholders. However, the general M&E policy and system (including the different instruments) is developed at the main office in Belgium and adapted at intervention level at a later stage.
The findings in table 4 show that with one single exception, all of the items relating to the analysis score relatively well for both relevance and effectiveness, with relevance scoring a little higher than effectiveness. More specifically the different scores can be described as follows.

Target groups are generally described well and unambiguously (items 1.1.1 through 1.1.3). However, in many cases the target groups are not clearly defined, which clearly affects the later evaluability. The (problem) situation is usually well described, and that is also the case for how the intervention relates to it, however without detailed descriptions. Although this indeed would not seem desirable from the reduced administrative requirements standpoint, it does constitute a handicap for evaluability.

As regards the description of the key stakeholders outside the target group (1.1.4), it must be noted that the difference between target groups and other stakeholders is not always clear, nor explicitly included in most propositions, not in the least because such information is not requested. From an evaluability point of view, a better description of the aggregate stakeholders would be required, as well as a more detailed description of the intermediate and final target groups, including their respective roles. At the same time one must consider the fact that some interventions are highly complex and involve a large number of stakeholders, which makes the underlying analysis quite difficult.

The integration of a gender analysis in the underlying analysis (1.1.5) is the lowest scoring item in this component. The weak scores come as no surprise and confirm the findings from the recently finished evaluation with respect to gender sensitivity of the Belgian development cooperation. The absence of a gender analysis could undermine the evaluation of an intervention, as possible reasons for failure of the intervention would not surface as a result of it. More specifically, since gender relations strongly affect the attitude of individuals, these relations often affect the implementation and results of interventions. Disregarding the influence of gender often has negative consequences for the quality of evaluations and the added suggestions for improvement. In this context it is noteworthy that many interventions pay little attention to social differentiation in the target group (not just from a gender standpoint) and that target groups are often treated as a homogenous group. This forms a handicap for the evaluation of relevance and effectiveness, since social differentiation can make that a project is either more or less relevant for certain (sub) target groups and that projects based on insufficiently recognised social mechanisms can score differently in terms of effectiveness as regards those different target groups. From a broader perspective this enhances the tendency to not or insufficiently look at the effects of intervention in terms of equality and equity, even though these principles take a central position in development cooperation.

The relatively high scores (within the 20% best scoring items) for the description of the relation between the underlying analysis and the intervention objectives (1.1.6) must be put into perspective. It would appear, particularly in the case of a second or third phase of an intervention or of programmes, that the underlying analysis is being written for the purpose of objectives that were determined earlier. This shows, among other things, from the fact that alternative strategies are not taken into consideration, but that the analysis is concentrated on one specific strategy which is clear from the start. Referring to the classic PCM cycle, we could state that in many cases, instead of performing a veritable identification in which different strategies are balanced against one another, one immediately jumps to the formulation. From the standpoint of evaluability, this implies additional challenges for evaluating the relevance, which in such cases must be

---


24 As an aside, it can be noted that equality/equity are not incorporated in the five classic DAC criteria which most evaluations currently focus on.

25 NGAs in particular have a tendency to build their programmes on experiences from the past. Although that is not a negative thing as such, it does entail a few dangers, as shown here.
The relation between the rationale of the intervention and the sector policy of the partner country (1.1.7) is clear in the majority of interventions, which translates into the highest score within this component and a place among the 20% best scoring items (see also table A3 2 of Appendix 7). The high score shows that the stakeholders responsible for the interventions are aware of the sector policy and seek to match the interventions to said policy, even if application (or applicability) of the policy is limited. It is quite likely that the increasing experience within the Belgian development cooperation of working with more general, long-term indicative programmes, which aim for more co-ordination and better synchronisation with the (sector) policy of the partner country, plays a role here. The exchange between the Belgian stakeholders and existing platforms in the partner countries may also have a positive effect.

3.1.2 The intervention logic and the Theory of Change

Under this component we examined in how far the intervention logic and the theory of change have been clearly described and up to which level (output, outcome, impact level). As indicated below throughout the analysis of the individual items, the quality of the intervention logic and the theory of change is an important factor in evaluability. Within our analytical framework, the attention for the specific level up to which intervention logic (as well as the crucial links, assumptions, internal risks, etc.) has been specified clearly and properly, has implications for the evaluability of the individual OECD/DAC criteria. In other words: intervention logic that is well specified up to the level of the outcomes, but not up to the impact level, improves evaluability of effectiveness but holds a considerable challenge for the evaluability of the impact. 8 items were analysed under this component in total (see table 5).

Of the eight examined components, intervention logic and theory of change has the third lowest evaluability index (see Table 3). As indicated below, the main bottleneck is how to specify the intervention logic and theory of change at the level of the impact. In other words: most interventions focus on the lower levels and on developing the implementation theory, whereas the higher levels and more specifically the impact level (which in our analysis coincides with the level of the general objectives from the logical framework) are often neglected. Nevertheless, effects at impact level (particularly the effects that arise directly from the outcomes) should be visible after closure of a three-year intervention (the term of many interventions in accordance with the funding cycle) and especially after a follow-up intervention of the same term.

The findings in table 5 show that for this component the levels effectiveness and efficiency are usually well-specified. A relatively large number of interventions achieve high scores for their theory of change which is logical and realistic up to the output (70%) and outcome (60%) level; this item (1.2.3) is one of the 20% best scoring items. Critical and crucial links relatively often get to be identified to both levels. Many interventions also indicate internal risks and external assumptions, although very often, these don’t get to be investigated further. Although the difference between the above-mentioned percentages for output and outcome are not very big, it is clear that in general, the output level (efficiency) is specified better than the outcome level (effectiveness).

26 At first sight this finding does not seem impressive, but it is nevertheless important since the majority of the best scoring items belong to dimension 3 (the context); see also tables A3 1 and A3 2 of Appendix 7 for more details.
27 These three levels are in accordance with the outputs (intermediate results), the specific objective and the general objective(s) of the logical framework. See also chapter 2.2 for how this study interpreted the five DAC criteria.
Table 5: Main results for intervention logic and theory of change

<table>
<thead>
<tr>
<th>Feature</th>
<th>Relevance</th>
<th>Effectiveness</th>
<th>Efficiency</th>
<th>Impact</th>
<th>Sustainability</th>
<th>Evaluability index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2 The intervention logic and theory of change</td>
<td>2,80</td>
<td>3,00</td>
<td>3,50</td>
<td>2,10</td>
<td>3,25</td>
<td>2,96</td>
</tr>
<tr>
<td>1.2.1 There is a clear and correct distinction between output, outcome and impact</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2.2 The theory of change from input to outcomes and the final impact has been worked out clearly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2.3 The theory of change is logical and realistic</td>
<td>4,15</td>
<td>4,40</td>
<td>3,20</td>
<td></td>
<td></td>
<td>3,92</td>
</tr>
<tr>
<td>1.2.4 Critical and crucial links in the chain have been identified and can be tested</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2.5 The theory of change contains concrete measures to safeguard the sustainability and the benefits of the intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2.6 Internal risks are clearly indicated and examined/estimated</td>
<td>3,00</td>
<td>3,10</td>
<td></td>
<td>2,80</td>
<td>2,97</td>
<td></td>
</tr>
<tr>
<td>1.2.7 External assumptions are clearly indicated and examined/estimated</td>
<td>2,80</td>
<td>3,10</td>
<td>3,20</td>
<td>2,15</td>
<td>2,90</td>
<td>2,83</td>
</tr>
<tr>
<td>1.2.8 The allocation of the provided means to the output is clear</td>
<td></td>
<td></td>
<td>3,64</td>
<td></td>
<td></td>
<td>3,64</td>
</tr>
</tbody>
</table>

The fact that a substantial part of the interventions specify crucial elements at the level of outcomes and particularly outputs, is a positive thing especially for learning at intervention level. Particularly for the directly involved intervention staff and the organisation responsible, it is important to acquire an insight into elements that are essential in the implementation of an intervention and therefore should be copied into future interventions, whereas other elements of a more accessory nature, could be left out. Especially in the case of innovative interventions where it is insufficiently clear what works and what doesn’t and which would be the best implementation methods, identification of the critical and crucial elements is essential\(^{28}\). Then again, identifying and testing external assumptions is important mainly for accountability, as it makes clear which external factors (could) affect the intervention.

As indicated before, the next step in the theory of change, the impact level, yields a considerably lower score. More precisely, the weakest elements within the intervention logic and the theory of change involve the impact criterion in particular. In a relatively large number of interventions the theory of change is not specified up to the impact level and those cases where it is, the impact level is usually at a considerable distance from the outcomes. The result is a missing middle, which makes it impossible to determine the contribution of an individual intervention to that impact\(^{29}\). The fact that a definition of critical and crucial links and of external assumptions between the outcome and impact


\(^{29}\) This can be solved, e.g. by inserting an extra level into the theory of change and working with intermediate and final outcomes to develop an extra step between outcomes and impact. We will get back to this in the recommendations.
level is often missing as well, enhances the effect (the score of this item (1.2.4) sits among the 20% lowest scores). It is clear that this undermines the feasibility to perform impact evaluations as well as their quality, since it is exactly the analysis of these elements which allows for differentiation between problems at implementation level, faults in the underlying theory of change and influences of other external factors that are outside the control of the intervention. Furthermore, particularly the identification and testing of critical and crucial factors can provide a better understanding of the reasons why an intervention generates impact or not. This partly undermines the learning/feedback function at the super-intervention level which is supposed to provide policy makers in particular (but also other stakeholders who are not directly involved in the implementation of specific interventions) with interesting views on outcome and impact. Furthermore, specifying a TOC up to the impact level can also provide insight into the implementation models that are most effective and efficient and provide the best chance for realising impact (without having to perform an evaluation of the impact each time).

The weak scores found for the impact criterion can perhaps be explained in part from the fact that – especially in the case of small interventions or organisations, but also in general – there are not many incentives to exceed the implementation level: specifying an intervention logic at implementation level is of far greater direct usefulness to the intervention/organisation itself and it is also the level over which the intervention/organisation has most control. On the other hand, the links to the higher levels are often unclear and less controllable, while identification of critical elements and crucial links at higher levels is harder and often necessitates a more in-depth sector and context analysis. Furthermore, until recently the funding bodies attached little importance to impact theory. This becomes clear from the fact that the formats of the project propositions in fact permit that this level receives little to no attention. The assessment format of DGD for Zuid actions does not pay any attention to impact either, possibly because funding bodies still attach more importance to accountability and because they realise that interventions (can) only have actual control over activities and outputs. As will be shown in this report, the absence of a fully fleshed out theory of change also affects the M&E system (both on paper and in practice) and this puts considerable pressure on evaluability in particular.

Although the way in which the interventions have been specified has a negative effect on the evaluability of the impact, the scores for sustainability are somewhat more positive. Most interventions formulate concrete measures and even outputs to safeguard the sustainability of the benefits of the intervention. This reflects the increasing concern among organisations to retain the results of their intervention after termination of the interventions. However, in the assessment of internal risks and external assumptions, sustainability is taken into consideration to a much lesser extent: 28% of the interventions have indicated and examined internal risks related to sustainability, whereas only 18% have done so for external assumptions.

Finally, all the interventions provide information relating to the provided means and in only 10% of the interventions, the allocation of the provided means/costs to the outputs is not clear at all. The good score for this item has a positive effect on and is, in fact, of crucial importance for the evaluability of the efficiency; it is also, in part, a result of the increasing attention for this criterion in the financial formats imposed by DGD.
3.1.3 The proposed M&E system

Under this component we examined in how far the intervention design provides information about the M&E system. The quality of the (proposed) M&E system is obviously an important factor of evaluability: the better it is developed, the better it is suited to generate information that will benefit evaluability.

An analysis of the nine items under this component has given us an insight into the proposed M&E system. In this context it is important to consider the fact that insight into the proposed M&E system does not say anything about the (quality of the) actual application of this system (such actual application will be dealt with in part 2.2).

Table 6 contains a summary of the main findings in table format, in keeping with the presentation for the previous two components

<table>
<thead>
<tr>
<th>Table 6: Main results with respect to the proposed M&amp;E system</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relevance</strong></td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>1.3 The proposed M&amp;E system</td>
</tr>
<tr>
<td>1.3.1 The main targeted results of the intervention have been operationalised in a good manner</td>
</tr>
<tr>
<td>1.3.2 Where necessary/relevant the indicators are gender-specific and/or have been split into other relevant parameters</td>
</tr>
<tr>
<td>1.3.3 The proposed M&amp;E system contains a consistent translation of the underlying intervention logic and theory of change</td>
</tr>
<tr>
<td>1.3.4 The working method for following up and evaluating how the intervention results and their sustainability were achieved is clearly described</td>
</tr>
<tr>
<td>1.3.5 The working method for following up the assumptions is clearly described</td>
</tr>
<tr>
<td>1.3.6 The working method for following up the internal risks is clearly described</td>
</tr>
<tr>
<td>1.3.7 the personal and financial resources of the M&amp;E system are clearly described</td>
</tr>
<tr>
<td>1.3.8 The proposed MIS allows for an allocation of expenses to specific outputs and intervention components</td>
</tr>
<tr>
<td>1.3.9 The way in which the M&amp;E system of the intervention matches/relates to the national/local M&amp;E system is clearly described.</td>
</tr>
</tbody>
</table>
The proposed M&E system is the component with the lowest evaluability index of the eight examined components and for the five DAC evaluation criteria. Seven (!) of the nine items of this component sit in the group of the 20% weakest scoring items for at least one criterion and often for all of the criteria (see also table A3 3 of Appendix 7). From the standpoint of theoretical evaluability, this need not be a problem as such. Ex ante, particularly the quality of the underlying analysis and that of the intervention logic and theory of change are of importance for the evaluability. The quality of the M&E system is important mainly from the viewpoint of the practical evaluability, but in that context it is mainly the M&E practice that matters, not so much the M&E system on paper as it has been proposed in the intervention proposition. However, as we will see later on, for four criteria there is a strong correlation between the quality of the M&E design as proposed in the intervention proposition and the final M&E practice (see the discussion of component 2.2 below). From this perspective, the (quality of the) specification of an M&E system in the start-up phase is a good indicator of the later practice.

There are several explanations for the weak score of this component. First of all, the requirements of the donor as regards the description of the M&E system are not strongly detailed in the intervention proposal. Apparently other aspects are – justifiably or not – considered more important. This may be a reason why intervention promoters make but modest investments in this component. This hypothesis finds support in the fact that the scores for the actual M&E practice are considerably higher (about 20%) (see table 3). Another reason could be that promoters seldom have the habit of describing their M&E system in detail. Furthermore, from a mental point of view one does not pay any attention to M&E until implementation has started. As we will see later on, a few interventions only start defining and developing their M&E system at the start. Another important point is that at the moment of the study (spring 2015) several organisations had only recently developed an M&E strategy (or were working on it), which means that operationalisation in the field is lagging behind.

Looking at the five DAC evaluation criteria, it comes as no surprise that 'impact' has the lowest score and that effectiveness and efficiency have the highest scores, while it must be noted that the score for these two criteria is still below the medium score of 3. The relatively high scores for effectiveness and efficiency can obviously be explained from the fact that M&E systems primarily aim at achieving the specific objective and particularly, the outputs to be realised, plus – in connection therewith – from the fact that most organisations' M&E policies attach greater importance to monitoring than to evaluation. The lower level of attention for impact in the intervention logic and theory of change (see under component 1.2 above) also plays an important role. That the scores are under par also has to do with the fact that in some cases, the basic skills for correct formulation of proper indicators apparently are not available.30

The statement that M&E systems mainly target the outputs and underlying activities and means is also confirmed by the relatively high score, at the efficiency level, for the operationalisation of the desired results (1.3.1). On the other hand, the score for impact is weak here, too, mainly because in many cases, impact is not given any attention at all. In that sense the proposed M&E system constitutes a pre-reflection of the M&E practice (see chapter 2.2 below where we will find that M&E systems are reasonably well-developed at an operational level (monitoring in the narrow sense), but that their quality and scope diminishes as one moves up in the ends-means chain where a broader focus is required).

Also with respect to the transformation – in the M&E system – of the underlying intervention logic and theory of change (item 1.3.3), the scores are highest in the fields of efficiency and effectiveness. Up to a certain level M&E systems also appear to pay

30 In particular, organisations take little account of the requirement that indicators must be ‘specific’ (the “S” in SMART): the indicator must particularly and in a ‘specific’ way refer to the result which is to be achieved via the objective (output, outcome, impact).
attention to activities and outputs that should improve the sustainability of the interventions (item 1.3.4); the scores in question are not really good, but nevertheless indicate explicit attention for sustainability in many interventions. Here, too we can discover a strong parallel to the findings concerning sustainability in the analysis of the previous component (see the discussion on component 1.2 above).

The (very) low scores for the use of gender-specific indicators (and other factors of social differentiation) (1.3.2) and the follow-up of external assumptions (1.3.5) can be explained to some extent from the lack of attention for these points in the underlying analysis and the specification of the intervention logic and theory of change (see the discussion of components 1.1 and 1.2 above). Apparently here, too, the hypothesis holds ground that if these elements are not included right from the start (e.g. via gender-specific baseline data and risks), they do not get sufficient attention later on. On the other hand, the reverse situation is also true, as illustrated from the good practice in Box 2.

Box 2: Risk analysis in the PASAB II project (Caritas)

The project documents of PASAB II (Projet d’Appui à la Sécurité Alimentaire au Bugesera), an intervention of the NGO Caritas in Rwanda, include clear risks for each intended result which could affect achieving those results. The project also encompasses a risk analysis indicating, for each of the risks, the possible effects as well as actions and modifications to the programme which will be implemented if the risks actually materialise.

Besides the follow-up of the results in a format which provides room for adding qualitative comments, the risks are also monitored via a specific format which compares the expected risks with the actual situation on the terrain. These documents contain useful information that can also be used in evaluations, because they provide an insight into the underlying reasons for the success or failure of interventions.

The equally low score for the method to monitor the internal risks (1.3.6) can be explained, in our opinion, from the fact that few organisations are inclined to describe the risks in their intervention propositions (see also the discussion on component 1.2 above). This standpoint is understandable: the organisations know that their propositions will be screened thoroughly and they are therefore eager to avoid showing a weak position. There are also organisations who find that these risks are part of their internal affairs, are not directly connected to the quality and implementation of the programmes and therefore, need not be/are not shared. These arguments seem convincing enough if one compares the scores to the quality of the information on the practice of risk analysis, which shows a higher score (see also item 2.1.9). On the other hand, from the viewpoint of evaluability it is desirable and important that information is available on these internal risks, the way in which they are dealt with and how they will be monitored.

In general, a description of the personal and financial means available for the M&E system is either absent or very limited (1.3.7). Only 25% of the propositions contain substantial information on these matters. Oftentimes, planned evaluations, even the ones of considerable size, are not mentioned separately in the budget. It is assumed that this is another case of not finding such information of crucial importance. From an evaluability standpoint it implies limited insight into the importance attached ex ante to M&E.

The way in which the M&E system of the intervention relates to the national/local M&E system is not clearly described in most cases, as shows from the very low score for this item (1.3.9). The low score can be explained, in part, from the fact that some interventions are very small or address matters that are not integrated into M&E systems at a higher level. Other factors involved include the absence, the doubtful reliability or the non-functioning of such systems at a national and decentralised level (e.g. DRC), which make that they are not considered as adequate to comply with the upward
accountability towards the funding authority. Nevertheless it is also clear that too few interventions in fact check in how far their M&E system – and particularly the monitoring of key indicators for outcome and impact levels – could be in line with what happens at higher and decentralised levels (or partner organisations)\(^\text{31}\).

The highest score under this component was found for the item *the proposed MIS allows for allocation of expenses to specific outputs and intervention components* (3.1.8). From an evaluability viewpoint this is an important achievement, which has to do with the way in which the donor wants to have the budgets spent and with the importance attached to a properly specified budget in the proposition. The imposition of a format, which connects several categories of expenditure to outputs, results in important information for an efficiency analysis.

Last but not least it is important to note that few interventions actually have an M&E system. Although the propositions usually contain information on the components of such a system, they are not (or only partly) synchronised and combined into a coherent whole. For instance, very often the risk analysis (internal risks and external assumptions) contains important factors that are missing from the logical framework, and vice versa. The absence of an actual system extrapolates into the M&E practice.

### 3.1.4 The consistency and adaptation of the intervention logic and theory of change

For the evaluability of interventions it is essential that any changes in the intervention logic and in the underlying theory of change are indicated and modified *clearly* in the M&E system. It is equally important to indicate clearly why and how the changes have been implemented. Failure to clearly indicate and modify the changes in the M&E system results in lack of relevant information with respect to the actual situation, which in turn can lead to wrong conclusions about the quality of interventions (related to all OECD/DAC criteria) and the factors that affect the implementation and effects of interventions. If the M&E system is not modified, it also implies that part of the renewed intervention will not be monitored.

In about one quarter of the interventions, changes have been made to the intervention logic and the underlying theory of change during the implementation phase\(^\text{32}\). Most of these changes are limited to output and the underlying levels.

This component scores relatively well in comparison to the other components of dimensions 1 and 2 (see table 7). However, the scores for the three items show considerable variation: relatively high scores for indicating and substantiating any changes to the intervention logic and underlying theory of change (which in part are among the 20% best scorers), against relatively low scores for the availability of information on the vision and opinions about any changes of the main stakeholders.

---

\(^{31}\) The consequences of this flaw will be discussed in 2.2 of this chapter.

\(^{32}\) The limited number of interventions that we could analyse here, calls for caution with respect to the findings.
Table 7: Main results for consistency and adaptation of the intervention logic and theory of change

<table>
<thead>
<tr>
<th>1.4 Consistency and adaptation of the intervention logic and theory of change</th>
<th>Relevance</th>
<th>Effectiveness</th>
<th>Efficiency</th>
<th>Impact</th>
<th>Sustainability</th>
<th>Evaluability index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4.1 Any changes in intervention logic and underlying theory of change are clearly indicated and substantiated</td>
<td>4,27</td>
<td>4,00</td>
<td>3,76</td>
<td>4,20</td>
<td>4,17</td>
<td>4,03</td>
</tr>
<tr>
<td>1.4.2 Information on the vision and opinions of the main parties involved about any changes in intervention logic and theory of change is available</td>
<td>2,82</td>
<td>2,63</td>
<td>2,43</td>
<td>2,80</td>
<td>2,67</td>
<td>2,63</td>
</tr>
<tr>
<td>1.4.3 Any changes in intervention logic and theory of change are incorporated adequately into the M&amp;E system</td>
<td>3,18</td>
<td>3,25</td>
<td>3,10</td>
<td>3,00</td>
<td>3,33</td>
<td>3,17</td>
</tr>
</tbody>
</table>

The fact that changes are indicated in most interventions is a positive finding in itself and illustrates the quality of the management and, at the same time, the desire to document and substantiate changes. Although the score for adequately modifying changes in the M&E system are reasonable, the fact that a large part of the interventions where changes have been implemented (and indicated) have not extended them to the M&E system, is cause for concern. Weak individual scores for this item are often associated with weak scores for the entire M&E system: in cases where systems are weak, changes usually are not implemented into the M&E system.

3.2 Analysis and practice with respect to implementation and management of the intervention and the context

This part deals with two strongly detailed components: the (availability of) basic information (11 items) as regards the intervention implementation, and the M&E system in practice (12 items).

3.2.1 The availability of basic information about the intervention implementation

For this component we examined whether the basic information about the intervention implementation is available. This involves information which has to be available at the start of an intervention (e.g. intervention proposition; baseline) as well as information relating to the progress of the intervention. The availability of basic information concerning the intervention implementation is essential for evaluability, since without such information it is hard to make a comparison between the start, intermediate and end situation, which in turn makes it hard to indicate and hence evaluate the progress of the implementation and the effects of an intervention.
Table 8: Main results for the availability of basic information about the intervention implementation

<table>
<thead>
<tr>
<th>2.1. Availability of basic information about the intervention implementation</th>
<th>Relevance</th>
<th>Effectiveness</th>
<th>Efficiency</th>
<th>Impact</th>
<th>Sustainability</th>
<th>Evaluability index</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1.1. The basic documents are available</td>
<td>2,95</td>
<td>3,20</td>
<td>3,53</td>
<td>2,13</td>
<td>2,62</td>
<td>2,88</td>
</tr>
<tr>
<td>2.1.2. Baseline information about the target group is available</td>
<td>4,03</td>
<td>4,05</td>
<td>4,13</td>
<td>3,73</td>
<td>3,95</td>
<td>3,97</td>
</tr>
<tr>
<td>2.1.3. Baseline information for the indicators with respect to the counterfactual is available</td>
<td>3,50</td>
<td>3,90</td>
<td>2,45</td>
<td>2,84</td>
<td>3,18</td>
<td></td>
</tr>
<tr>
<td>2.1.4. Baseline information is available, and relevant indicators have been split gender-specifically and/or into other relevant parameters</td>
<td>1,28</td>
<td>1,28</td>
<td>1,44</td>
<td>1,29</td>
<td>1,32</td>
<td></td>
</tr>
<tr>
<td>2.1.5. The information to be expected on the progress of the realisation of the intervention objectives, is available</td>
<td>1,76</td>
<td>1,86</td>
<td>1,51</td>
<td>1,71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1.6. Information on the participation of the original target group is available</td>
<td>3,26</td>
<td>3,31</td>
<td>3,41</td>
<td>2,84</td>
<td>3,11</td>
<td>3,19</td>
</tr>
<tr>
<td>2.1.7. Information about data collection for the indicators is available</td>
<td>2,73</td>
<td>3,08</td>
<td>3,28</td>
<td>2,28</td>
<td>2,69</td>
<td>2,82</td>
</tr>
<tr>
<td>2.1.8. The collection of data in principle allows for reliable collection with respect to the indicators</td>
<td>3,00</td>
<td>3,12</td>
<td>3,47</td>
<td>2,52</td>
<td>3,00</td>
<td>3,04</td>
</tr>
<tr>
<td>2.1.9. The information related to the monitoring of internal risks is available and any consequences for the intervention logic and implementation have been indicated</td>
<td>2,40</td>
<td>2,75</td>
<td>2,13</td>
<td>2,86</td>
<td>2,54</td>
<td></td>
</tr>
<tr>
<td>2.1.10. The information related to the monitoring of assumptions is available and any consequences for the intervention logic and implementation have been indicated</td>
<td>2,21</td>
<td>2,37</td>
<td>2,60</td>
<td>2,20</td>
<td>2,45</td>
<td>2,37</td>
</tr>
<tr>
<td>2.1.11. The intervention expenditure is well-documented/recorded and can be allocated to the outputs</td>
<td>3,95</td>
<td>3,95</td>
<td></td>
<td></td>
<td></td>
<td>3,95</td>
</tr>
</tbody>
</table>

Of all of the 8 examined components, the component *Available basic information concerning the implementation of the intervention* has the second lowest score on the evaluability index (see table 3). This is somewhat surprising: from an M&E standpoint it may be assumed that 'having information available' would be a first step. As becomes clear from the discussion of the different items, the scores show considerable variation and particularly very low results for the availability of basic information regarding a 'counterfactual' (which particularly complicates the evaluability of impact) and for the availability of information that is split according to gender or other relevant parameters. A more nuanced finding is that nevertheless a lot of information is being collected and available, but that it is very much concentrated at the lower levels of the chain and hardly relates to elements *falling outside the intervention* (such as the counterfactual, the external assumptions, etc.) on the one hand, and to the data collection process in itself (sort of meta-information), on the other. This finding largely correlates to earlier findings for the components regarding the intervention logic (see 1.2) and the proposed M&E system (see 1.3), where it became clear that the focus is mainly on the operational and implementation level. The fact that neither a tradition nor clearly formulated requirements exist as regards the submission of information about the actual set-up of M&E (method of data collection and processing) could play a role here too.
As shown in table 8, the strongest item in this component is the ‘availability of basic documents’ (2.1.1)\(^{33}\). Basic documents include intervention propositions, the technical and financial file, baseline reports and progress reports. Up to the outcomes level these documents relatively often contain information on the progress of the realisation of the intervention objectives (2.1.5), making a comparison to the start phase in the process. This information is important for going one step beyond monitoring. However, for obvious reasons it is hardly available at the impact level. The strong score for the availability of the basic documents calls for some nuancing, as in some cases the information is available only in a language that some of the stakeholders do not master. Furthermore, the available documents are not always known to all of the stakeholders, which implies a limited ownership of the contents of these documents among these stakeholders. This affects the evaluability of an intervention because in such cases, the evaluation is based on information that is not known to all actors and it therefore depends on the specific stakeholders in the evaluation whether certain information will or will not be available\(^{34}\).

Two items show extremely low scores for the five DAC evaluation criteria: the availability of baseline information concerning the counterfactual (2.1.3) and the split into gender or other parameters of relevant indicators in the baseline (2.1.4); both items are high on the list of the 20% items with the weakest scores (see also table A3 3 of Appendix 7). The absence of disaggregated information is not surprising and can be explained from the absence of a gender analysis and disaggregated indicators in the initial M&E system (see above).

The counterfactual is an estimate of the situation which would occur if the intervention would not have taken place and is mainly important for a methodologically correct implementation of impact evaluations\(^{35}\). The bad scores for this item are in part caused by the fact that identification of a good counterfactual is not always easy, that it requires some methodological knowledge and that it involves practice not yet common in the Belgian development cooperation. The use of a counterfactual also appears to be interpreted in a limiting sense and downsized to the use of an RCT (Random Controlled Trial) type of control group. Taking this route is not always possible, or even desirable; but there are plenty of alternatives available that are often less complicated but apparently also less known\(^{36}\). Examples include the selection of a non-intervention group via 'matching'; selection of a similar group involved in a different intervention (to acquire an insight into differential impact); the use of statistical checks, generic checks and more complex types of pre and post comparisons, such as panels, time series or shadow checks\(^{37}\). A few interventions from the sample have used a counterfactual that was not complex and not expensive; a few others could identify a counterfactual relatively easily.

---

33 ... although in a few cases the parties involved had to be reminded several times before they finally submitted the information. In addition it is important to note that the quality of those documents varies considerably. This shows from the scores on the other items.

34 For this study it also meant that for a large number of interventions, the scores allocated for this item on the basis of the preliminary desk study had to be altered during the field mission. In other words, if an evaluator is unable to perform a preparatory field visit, this has important implications for the extent to which he/she will be able to prepare for the field mission and therefore indirectly affects the evaluability.

35 The term impact is used here in the meaning of the effects directly attributable to the intervention (causality) (see also footnote 14) and which therefore could be situated at the level of the outcomes as well as the higher situated links in the chain.


37 In the case of matching, a control group is sought which is similar to the intervention group in terms of the characteristics that are thought to influence the results of the intervention. In the case of statistical checks the logic is similar, but here the determining characteristics are not checked at the moment of the evaluation set-up and data collection, but at the moment of the data analysis. In other words, the point is not to find a similar group at the evaluation set-up, but to check for determining characteristics by adding them to a regression as control variables, which allows for fine-tuning of the influence of these variables. The disadvantage of the aforementioned counterfactuals is that they require collection of data from a group that is not involved in the intervention – something that is not always possible or desirable. In the case of generic checks one uses nationally available data about the outcome one seeks to achieve with a specific intervention. In other words,
Most interventions provide baseline information with respect to the (participation of the) target group (2.1.2 and 2.1.6), but this information is often incomplete and restricted to those effectively reached. Not in the least because the information with respect to the (definition of the) target group is usually of a general nature in the intervention proposition, said information often is quantitative and provides little qualitative insight into the situation of the target group. This is somewhat surprising as baseline studies often have been carried out, but they will provide a general view on the region or sector rather than an in-depth analysis of the specific characteristics of the target group (and a counterfactual), or of the factors that could affect the implementation and the results of the intervention. It is exactly this type of baseline analysis that allows for identifying parameters (such as gender) which could affect the implementation and the results of an intervention, and had therefore best be included in the M&E system as they provide crucial information that seriously influences the quality and exhaustiveness of the evaluation.

The presence of the broadest possible disaggregated baseline information is essential for the evaluability of interventions as it provides a means to compare the start and end situation. Specific information concerning the effectively reached target group compared to the originally defined target group is, in turn, essential to check whether the intervention reaches its original target group. This information is a first step in the analysis of the way in which the intervention does or fails to reach its target group and allows for evaluating whether the intervention was designed in such a way that the preliminary target group could indeed be reached effectively. Furthermore, the absence of information on drop-outs can easily lead to overestimating the effects of interventions.

The scores for the items relating to the monitoring of internal risks and assumptions (2.1.9 and 2.1.10) are moderate to weak for all of the criteria, with both of them belonging to the 20% weakest scoring items, whereby in general the monitoring of internal risks has a slightly better score than the monitoring of external assumptions. Furthermore, the output level scores slightly better than the outcome level, and the outcome scores better than the impact level. This finding is not surprising and is connected to the earlier mentioned focus on the implementation level and on elements directly connected to the intervention.

Most interventions have information on data collection for the indicators available up to the outcome level (2.1.7), but for very few of the interventions that information is complete and includes details about what data is being collected, by whom, how frequently and with what type of coverage, and what are the methods of data collection (see box 3 for an example of good practice). Furthermore, information on data collection is often available in a very fragmented form, both spread out over a few documents and spread out over several chapters in the same document. Such information about the data collection process is important for the evaluability of interventions, because the existing information often is a first data source for an evaluation and consequently, it is
essential to acquire an insight into the quality of the information. With respect to the reliability of the data collection (2.1.8) the most important conditions for reliable data collection (large enough sample; triangulation; proper frequency; independent data collector) are usually sufficient for evaluations of effectiveness and efficiency, but to a (much) lesser extent for evaluations of impact, sustainability and relevance.\(^{38}\)

**Box 3: Clear baseline information on the functioning of the M&E system at PROTOS (programme in Rwanda)**

The NGO Protos has developed a document called Scenario for monitoring indicators, which clearly indicates the persons and organisations involved in the data collection for each indicator. It also includes who is going to analyse the data. Furthermore, these scenarios describe the required method, moment and frequency for collecting the data for the indicator involved. They also mention the training provided – if any – for a specific method or tool. This kind of monitoring scenarios can be useful especially to maintain a general overview when several partners are responsible for specific components of the data collection and analysis, but also for providing an external evaluator with an overview of responsibilities and procedures of data collection and analysis.

### 3.2.2 The M&E system in practice

Under this component, the study examined how the M&E system works (or has worked) in practice. 12 items were defined which partly coincide with the items under 1.3 (*the proposed M&E system*). The quality of the proposed M&E system combined with the quality of its application is probably the component which most affects the evaluability of an intervention. Weaknesses in other components can, to a certain extent, be compensated by a properly developed and functioning M&E system. Conversely, the practical evaluability of an intervention will yield a low score if the M&E system functions badly, even if one has a good quality intervention design.

For all of the criteria, the *actual* M&E practice yields a substantially higher (more than 20%) score than the M&E system *proposed* in the project plan (see table 3 and point 1.3 of part 1 of this chapter). This confirms the earlier formulated hypothesis that for several reasons, the presentation of the M&E system in the project proposition receives relatively little attention (see 1.3). However, the big difference in scores is also an illustration of the importance attached to M&E in the *implementation* of the project and holds positive implications for the evaluability. The M&E, and particularly monitoring, is often better embedded in practice than can be decided from the project proposition. That is obviously a good thing, not only from a perspective of evaluability, but especially because (good) M&E forms an important part of project management and can make a substantial contribution to better interventions.

On the other hand, there is a strong correlation between the scores for quality of the M&E design (as included in the project proposition) and the quality of the M&E practice, and with the exception of impact, this applies to all of the evaluation criteria.\(^{39}\) The correlation is most outspoken for effectiveness and sustainability, which suggests that proper integration of these criteria in the M&E system from the start is of particular importance. Despite the existence of a strong correlation, the 40 studied interventions did contain examples of low-quality M&E systems in the project proposition which were corrected later on, for instance by focusing strongly on the development of an M&E system in the first quarter of the project implementation. The advantage of such an approach is that it provides a better insurance of safeguarding the ownership of the system and that it is easier to involve other stakeholders in the development and application of the system. Furthermore, a strong correlation does not necessarily imply

\(^{38}\) This item could be scored for no more than 24 interventions.

\(^{39}\) This probably has to do with the fact that impact receives little explicit attention in M&E systems – both in the proposition and in practice – which also reduces the chance of matching the proposition and the practice.
causality (from paper to practice): as shown below it is possible for the M&E practice to also affect the M&E proposition on paper – which is clearly possible in interventions that succeed previous interventions.

Table 9: The M&E system in practice

<table>
<thead>
<tr>
<th>Component</th>
<th>Relevance</th>
<th>Effectiveness</th>
<th>Efficiency</th>
<th>Impact</th>
<th>Sustainability</th>
<th>Evaluability index</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2 The M&amp;E system in practice</td>
<td>3,03</td>
<td>3,30</td>
<td>3,50</td>
<td>2,53</td>
<td>2,85</td>
<td>3,04</td>
</tr>
<tr>
<td>2.2.1 The vision on and the role of monitoring and evaluations (M&amp;E and independent evaluations) are clear</td>
<td></td>
<td>3,23</td>
<td>3,31</td>
<td>2,90</td>
<td>2,73</td>
<td>3,05</td>
</tr>
<tr>
<td>2.2.2 Sufficient time, means and personnel have been provided for the M&amp;E system to work adequately</td>
<td>3,67</td>
<td>3,87</td>
<td>4,15</td>
<td>3,21</td>
<td>3,63</td>
<td>3,71</td>
</tr>
<tr>
<td>2.2.3 The responsibilities and procedures for collecting and analysing M&amp;E data have been clearly defined</td>
<td>3,50</td>
<td>3,85</td>
<td>4,08</td>
<td>2,64</td>
<td>3,41</td>
<td>3,50</td>
</tr>
<tr>
<td>2.2.4 The responsibilities and procedures for decision-making based on the analysis of the M&amp;E data have been clearly defined</td>
<td>3,62</td>
<td>3,75</td>
<td>4,03</td>
<td>2,95</td>
<td>3,56</td>
<td>3,58</td>
</tr>
<tr>
<td>2.2.4bis The main parties involved agreed on the proposed M&amp;E system (including their role in it)</td>
<td>4,18</td>
<td>4,20</td>
<td>4,33</td>
<td>4,23</td>
<td>4,18</td>
<td>4,22</td>
</tr>
<tr>
<td>2.2.5 The staff responsible for M&amp;E are competent and independent</td>
<td>2,95</td>
<td>3,11</td>
<td>3,32</td>
<td>2,73</td>
<td>3,00</td>
<td>3,02</td>
</tr>
<tr>
<td>2.2.6 The M&amp;E system of the intervention is fine-tuned to the national/local M&amp;E system</td>
<td>2,59</td>
<td>2,88</td>
<td>3,12</td>
<td>2,52</td>
<td>2,64</td>
<td>2,75</td>
</tr>
<tr>
<td>2.2.7 There is an internal drive for strategic steering and learning</td>
<td>3,72</td>
<td>3,80</td>
<td>3,90</td>
<td>3,70</td>
<td>3,72</td>
<td>3,77</td>
</tr>
<tr>
<td>2.2.8 The results from M&amp;E are used for learning</td>
<td>3,05</td>
<td>3,08</td>
<td>3,30</td>
<td>2,93</td>
<td>3,00</td>
<td>3,07</td>
</tr>
<tr>
<td>2.2.9 The results from M&amp;E are used for accountability</td>
<td>3,23</td>
<td>3,40</td>
<td>3,63</td>
<td>2,95</td>
<td>3,08</td>
<td>3,26</td>
</tr>
<tr>
<td>2.2.10 Good quality evaluations and/or studies providing useful information have been carried out</td>
<td>2,67</td>
<td>2,78</td>
<td>2,75</td>
<td>2,22</td>
<td>2,58</td>
<td>2,61</td>
</tr>
<tr>
<td>2.2.11 The M&amp;E system is regularly checked for quality and adapted as necessary</td>
<td>2,74</td>
<td>2,90</td>
<td>3,00</td>
<td>2,80</td>
<td>2,85</td>
<td>2,86</td>
</tr>
</tbody>
</table>

On the other hand, this component yields but an average score in comparison to the seven other components we studied (table 3), which shows that there is room for improvement. This seems quite normal: the conversations with the project operators showed us that in many cases, the existing M&E systems were introduced or given shape quite recently. In many interventions it involves a process that is not yet complete, but where there clearly is an intention to refine both the structure and the implementation of the system. Said refinement initially involves creating an actual system from the different components that often were already present, but not yet fine-tuned to one another. Another remarkable thing was that many interventions display good practices as regards M&E, but fail in terms of reporting and oftentimes don’t show any link to other M&E practices. From the perspective of evaluability it therefore may be assumed that it will increase in the future via a further development of the M&E practice.

Furthermore, it seems that monitoring at the operational level is already being accomplished: the use of the resources, the monitoring of the activities and the progress

---

40 This item was initially included under dimension 1, but it was moved to component 2.2 during the analysis. For practical reasons the original numbers were maintained.

41 The data in this item involves only 27 of the 40 examined interventions. For the other 13 interventions we could not expect that an evaluation or study had already taken place.
in the area of the outputs have in many cases become an intrinsic part of the implementation practice. Nevertheless the evaluative component of the M&E system, and particularly the level of impact and sustainability, seems to be getting less attention. The possible causes are discussed below.

When comparing the scores of the five DAC criteria, a pattern emerges that is similar to that of the other components, particularly the lowest score at the impact level and the highest score at efficiency level. Equally remarkable is that the scores show a same ranking as for the description of the M&E system in the project proposition. In other words, the description of the M&E system, with all its imperfections, is the pre-reflection of the later practice. The low scores for impact, but also for sustainability, imply that the challenges to include these two criteria in evaluations already intrinsically (ex ante) present, have become even greater. On the other hand, effectiveness and particularly efficiency show quite good scores, which is a positive thing from the evaluability perspective, all the more since these criteria are the centre point for many evaluations.

Our analysis shows that the vision on (the role of) monitoring and evaluation is relatively under-developed, which is quite a surprise in light of the developing M&E practice (2.2.1), although we did find a few interesting exceptions in which policy changes give rise to the development of an M&E approach and an M&E policy and are supported with adequate human resources (see box 4). It comes as no surprise that the scores for impact and sustainability are lower in this context. What is surprising is that the difference is considerable: apparently the two criteria are often beyond the horizon when it comes to developing an M&E policy. If we dig a little deeper, we find yet a few other reasons. It is highly likely that up to a certain extent, the policy follows the practice instead of the other way around, and that that practice has already taken firm shape in an operational context. As such it would be understandable that in policy documents, monitoring if often better developed than evaluation. Furthermore it is also clear that vision and policy formation on monitoring and evaluation are often directed by the (Belgian) responsible organisation, but that it takes time to translate the policy into interventions, one of the reasons being that a ‘new’ or different policy can only be applied after consultation with the partners. In many cases this process has not yet terminated, which is understandable given the relatively recent attention surge for M&E. A consequence of the above is that in general the evaluability for effectiveness and efficiency is at a good level, which is also the case for relevance – be it for different reasons – but that this does not apply as much for impact and sustainability.

**Box 4: How a modified policy can lead to a better M&E approach**

Over the past few years, the strategy of Artsen Zonder Vakantie (Doctors without vacation) has seen substantial change. Whereas in the past the focus was on sending out Belgian medical staff to provide specialist care on site, today the demand from the partner organisations defines how medical support is being organised. This fundamental change has been extended to the M&E instruments. The organisation has developed a broad range of M&E instruments which should allow for better monitoring of the effects of the programme. At the centre is an instrument to measure the realisations in the area of capacity enhancement, which is based on the changes as defined/desired by the partner and includes the definition and monitoring of progress markers. Furthermore, the quality and effects of the missions of the Belgian medics are measured systematically via, e.g., a mission report jointly prepared by medic and partner according to a standardised format, and a specific report prepared by each separate party; these documents are combined into a synthesis memorandum at a later stage. All stakeholders receive feedback based on this synthesis report.

Table 9 shows the remarkably high score as regards the level of agreement among the main stakeholders about their role in the M&E system (2.2.4bis). This score calls for some nuancing since M&E is often the task of the project team with little (expected or requested) input from the other key stakeholders. Problems apparently arise only in more complex interventions or organisations, for instance when higher levels impose
demands as regards data collection, processing and reporting that are perceived as excessive by the lower levels.

The study also shows that there are usually sufficient means (time, personnel, funds) for M&E, although the impact level shows lower scores here for understandable reasons (2.2.2). From an evaluability perspective this is a positive finding, which does however require some nuance. Discussions with project teams revealed that although there was sufficient time to include ‘basic’ M&E functions such as collecting and processing data, time to discuss and analyse them in-depth often lacked. In most interventions M&E is not specifically assigned to a specific person or department, but rather taken up by several team members who are usually individually responsible for collecting data on ‘their’ output or the relevant part of it, and hence are limited to the operational level (inputs - activities - outputs). The focus on the operational level implies that the project teams concentrate on the ‘here and now’ and that the processes contain very few moments to take a step back and view things from a distance; this can also be a reason for the relatively low score for (regularly) reassessing the functioning of the M&E systems (2.2.11). Another reason why means are considered as sufficient is the relatively limited attention paid to evaluation (in the full meaning of the word); this will be described in more detail in part 3 of this chapter.

Responsibilities and procedures for the collection and analysis of M&E data and decision-making in this area are well established, with the exception of the impact level, which is often not included (2.2.3 and 2.2.4). Monitoring (and evaluation) at higher levels often simply does not form part of the M&E tasks. Such analysis is considered a collective responsibility and consequently, nobody feels specifically concerned. Although healthy peer pressure exists, everyone appears to take responsibility ‘only’ for their own part of the project implementation. The monitoring of general intervention effects and most certainly of impact are above that and moreover, require a more specific and demanding approach in terms of data collection and data analysis, which in turn require more specific methodological knowledge and more funds. In the development of many M&E systems, this is not really included and consequently, no responsibilities are assigned either. Furthermore, the information gathered through impact analysis often supersedes the specific intervention level, which makes that interventions/organisations often perceive it as less useful and directly usable, therefore not as a priority. Although such information is very useful at a higher (policy) level, it has features of a public good, as it were, which partly explains the deficient investment in impact and sustainability evaluations by specific interventions.

As an aside we would like to note that the desirability and option of involving other stakeholders in M&E seldom is examined thoroughly. As a result some opportunities go to waste in the area of learning, but also in terms of the sustainability of M&E systems and the possibility to spread the burden and responsibility for M&E across the different stakeholders.

Competence and independence of personnel with responsibility for M&E show lower scores than most of the other items in this component (2.2.5). A first explanation is that in small interventions, it is difficult to arrange really independent M&E, for instance by entrusting it to a separate service or person. This is an option for bigger interventions or programmes only (see box 5). As far as competence is concerned, there are few indications of specific expertise or investments in specific education.

So, most of the staff working on M&E are trained on the job. Generally this does not pose many problems, because (1) there is plenty of exchange and a good learning culture in the teams (see below); (2) the M&E task is limited largely to the operational level, which does not pose a lot of methodological challenges; (3) many interventions have developed good standard tools for important activities (e.g. education). However, it

---

42 In some cases local stakeholders must also meet monitoring requirements which are imposed by different stakeholders and often are not synchronised.
is clear that M&E – and particularly evaluation – is more demanding at higher levels and that untrained project staff cannot always meet the requirements. We can add that the methodological challenges at these levels are not recognised sufficiently and that in fact the outcome and impact levels within monitoring receive little attention. Furthermore we noticed that in many cases the quality and continuity of M&E systems suffer from staff turnover and internal changes. Especially when the M&E system is not yet fully formalised this can cause problems, which is an additional illustration of the importance of a more specified description of an M&E system.

**Box 5: The development of a cross-intervention M&E approach (BTC Katanga)**

The TFD of the EDUKAT project (carried out by BTC in the DRC) contains clear indications with respect to the set-up and organisation of an M&E system, which are based on the general policy and instructions developed by BTC in this regard. In this context, EDUKAT has taken the exceptional initiative to develop a work plan (Plan de travail baseline) which seeks to bundle the different M&E activities in a coherent whole. The work plan was developed after a reshuffle of the initial logical framework (which was considered quite difficult and not totally coherent) and defines the steps for developing a baseline, including a follow-up matrix (derived from the reshuffled logical framework), a plan for risk management and the effective application of the follow-up mechanisms. Implementing the plan is the responsibility of a programme manager who sees to the transversal management of M&E in different bilateral projects carried out by BTC in Katanga. The choice to assign one specifically responsible person was made from the desire to guarantee learning and continuity of the M&E system within a long-term (12 years) engagement. In this way it must also be possible to support a process of institutional strengthening via involvement of local structures in the data collection and data analysis and the follow-up of the M&E results.

The results of our analysis as regards the use of M&E for learning are quite good (2.2.7), with the exception of those for impact. Based on the strong orientation of the M&E at the operational level it is evident that the score is good in terms of efficiency. The use of the M&E results for accountability yields a somewhat higher score than the use for learning purposes (2.2.9). This may be explained from the fact that for accountability, the collected data as such largely suffices, whereas learning requires further analysis, reflection and internalisation, most of which – as indicated above – demands more time than is available. On the other hand, accountability is often interpreted in a very narrow way and is in fact limited to accountability towards the donor and, in some cases, indirectly towards the broader rank and file, for instance via websites. Very seldom we see conscious accountability towards key stakeholders such as target groups. We also found that relatively little feedback (about the data collection and the findings based on the analysed data) is given to local key stakeholders (often also the target group) involved in data collection processes. Such feedback is essential to avoid that local data collection becomes a ritual process with a negative effect on the reliability of the data – and hence of the evaluability – and, in the longer term, on the sustainability of the M&E system. This need for feedback applies not only to the relation between local intervention managers and the key stakeholders (including target groups), but also to the relation between the Belgian organisation and local intervention managers, and those of DGD towards the indirect stakeholders and BTC.

However, many interventions do show a strong drive for strategic direction and learning, which also translates into the implementation practice (2.2.7). This creates a framework which nonetheless compensates somewhat for imperfections of the M&E system (e.g. in the field of competences and independence and the limited attention for the higher levels in the ends-means chain). It also creates a project culture which is open to criticism and questions, which is an important ‘transversal’ advantage for the evaluability.

We have noted earlier that evaluation practice gets (too) little attention in the interventions. This is also illustrated by the weak score (it is the item with the lowest score under this component for all criteria; for 3 criteria this item is also on the list of the
weakest scoring items) with respect to the track record of the examined interventions in this area (2.2.10)\(^3\). We found several explanations for this that will be discussed in part 3. For now it suffices to refer to the strong focus on monitoring in many interventions, which is often rightly or wrongly based on the assumption that (external) evaluation does not have any added value. However, there are also examples of good practices in which the set-up and implementation of evaluations have been largely determined ex ante and applied over a longer period, thus ensuring that they can provide considerable added value (see box 6).

**Box 6: Development of a clear evaluation policy and practice**

As an NGA active in the field of development education, ITECO has developed a clear policy in the area of planning, monitoring and evaluation. As far as evaluation of its educational actions is concerned, ITECO has developed a standard approach which comprises three complementary dimensions that concern different links of the underlying theory of change: the evaluation of everything learnt during the education; the application (by the participants) of what has been learnt; and the effects of the education (by bringing the new skills into practice) at the level of the organisation and the environment of the participants. The results of the analysis allow ITECO to supersede the pedagogical domain and create an image of its contributions to the desired transformation. At the level of the programme co-funded by the Belgian authorities, ITECO receives support from an expert on education and training for refining the organisation's M&E instruments and defining the appropriate indicators. One of the results of this process was a high-quality logical framework with a clear baseline and annual targets, which should warrant proper monitoring of the progress that has been made and form an ideal basis for evaluations.

A last issue concerns the synchronisation of project M&E systems to national and decentralised systems (2.2.6). Not totally surprising the scores are quite low, except for interventions in Rwanda (see chapter 4.1). In many interventions, particularly those outside the bilateral channel, this issue is hardly addressed at all. As a result of the relation this is different in bilateral interventions, but other practical problems are often at play there, such as the weakness or absence of M&E systems at a national or decentralised level.

Incidentally, not one single project seems to have the ambition to improve existing M&E systems and make them more sustainable. Expecting that from interventions may be putting the bar too high. Nevertheless, each project should be checking whether its M&E system (or some of the key indicators) could be aligned to national, local or partner systems. After all, it is clear that the absence of a link to national or decentralised systems has important implications for evaluability, particularly if one wants to evaluate ex post and in that process, seeks to check, for instance, the sustainability of the benefits realised by the project.

\(^3\) ...but we must note that this item could be given a score for only 27 interventions.
3.3 The evaluation context

To check the effect of the context on the evaluability of the interventions, we made the assumption that an independent external evaluation would have to be carried out for each intervention. Furthermore, we made the choice to exclude practical context elements, such as the situation in terms of safety, infrastructure, etc. from the analysis, because they are highly specific for each intervention and hard to handle in the scope of this general study. However, it cannot be neglected that in reality, these types of elements are of great importance in deciding the evaluability. For this reason it was decided to mention them explicitly in the analytical framework.

3.3.1 The attitude of the key stakeholders

The attitude of the key stakeholders, such as the ones closely involved in the implementation of the intervention, but also the local authorities, donor(s) and other organisations active in the region, is of great importance to the evaluability of interventions. A negative attitude makes the work of the evaluators difficult under all circumstances, even if the other conditions, such as a good project plan or a properly functioning M&E system, have been met to a large extent.

Table 10: Findings with respect to the attitude of key stakeholders

<table>
<thead>
<tr>
<th>Sub-section</th>
<th>Relevance</th>
<th>Effectiveness</th>
<th>Efficiency</th>
<th>Impact</th>
<th>Sustainability</th>
<th>Evaluability index</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1. Attitude of the key stakeholders</td>
<td>3.68</td>
<td>3.70</td>
<td>3.70</td>
<td>3.58</td>
<td>3.63</td>
<td>3.66</td>
</tr>
<tr>
<td>3.1.1 The main users of the evaluation and their expectations/importance as</td>
<td>2.36</td>
<td>2.36</td>
<td>2.36</td>
<td>2.27</td>
<td>2.27</td>
<td>2.33</td>
</tr>
<tr>
<td>regards the evaluation have been clearly defined 44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1.2 The key stakeholders request (or at least are interested in) an</td>
<td>3.45</td>
<td>3.45</td>
<td>3.45</td>
<td>3.40</td>
<td>3.45</td>
<td>3.44</td>
</tr>
<tr>
<td>evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1.3 The expectations of the individual key stakeholders as regards the</td>
<td>4.11</td>
<td>4.11</td>
<td>4.19</td>
<td>4.11</td>
<td>4.16</td>
<td>4.14</td>
</tr>
<tr>
<td>evaluation (process and results) are compatible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1.4 The expectations of the key stakeholders as regards the evaluation</td>
<td>4.32</td>
<td>4.23</td>
<td>4.49</td>
<td>4.11</td>
<td>4.32</td>
<td>4.29</td>
</tr>
<tr>
<td>are realistic (in relation to the available resources)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1.5 The main users were/will be involved in the evaluation process</td>
<td>3.53</td>
<td>3.52</td>
<td>3.58</td>
<td>3.55</td>
<td>3.55</td>
<td>3.55</td>
</tr>
<tr>
<td>3.1.6 The main stakeholders (inc. target group) were/will be involved in the</td>
<td>3.62</td>
<td>3.62</td>
<td>3.62</td>
<td>3.56</td>
<td>3.56</td>
<td>3.59</td>
</tr>
<tr>
<td>evaluation process5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1.7 The relations between the key stakeholders are ‘healthy’</td>
<td>4.26</td>
<td>4.24</td>
<td>4.24</td>
<td>4.22</td>
<td>4.22</td>
<td>4.24</td>
</tr>
<tr>
<td>3.1.8 It is possible to contact all of the key stakeholders without running</td>
<td>4.63</td>
<td>4.64</td>
<td>4.64</td>
<td>4.63</td>
<td>4.63</td>
<td>4.64</td>
</tr>
<tr>
<td>the risk of mutual influencing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1.9 All of the key stakeholders have a positive attitude with respect to</td>
<td>4.18</td>
<td>4.18</td>
<td>4.18</td>
<td>4.18</td>
<td>4.21</td>
<td>4.18</td>
</tr>
<tr>
<td>independent evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

44 This item could be analysed for only 22 interventions; the other interventions were started too recently to allow for external evaluation.

45 An analysis of this item was useful and carried out only for 26 interventions.
As shows from table 10, positive scores dominate. The component involved yields a generally high score with five items on the list of 20% best scoring items (see also table A3 2 of Appendix 7)\(^{46}\). Another remarkable point is that the differences between the five evaluation criteria are minimal (however with impact and sustainability again showing the lowest scores), which can be explained from the consideration that the attitude of the key stakeholders often concerns ‘independent evaluation’ in its entirety. It is clear that a good score for the attitude of the key stakeholders is a positive finding in the light of evaluable. A positive attitude not only involves the evaluation process, but often is also an indication of openness and willingness to questions and criticism – also important factors for a good intervention implementation.

The extent to which the most important users of the evaluation and their expectations and interests as regards the evaluation have been clearly defined’ (3.1.1) is the item with the lowest score. There are various explanations for this finding. First and as indicated earlier, the policy and practice with respect to (independent) evaluation at project level are not well-developed (contrary to monitoring). In fact, at a project level evaluation gets little (explicit) attention and as a result, the same applies to the different parameters relating to evaluation: the possible objectives of the evaluation (accountability; learning; policy support) are hardly weighed against each other, hence neither are the choices arising thereof as regards the most important evaluation criteria and the possible (future) users of evaluation results (which don’t necessarily coincide with the key stakeholders in the implementation of the intervention\(^{47}\)). These points receive some attention in the preparatory phase of the evaluation, but often are insufficiently specified, which shows for instance from the analysis of the reference terms for evaluations. The absence of a clear effort (and clear choices) for this item affects the evaluable: failure to properly identify the expectations/interests of the key users and/or to make choices as regards the evaluation objectives inevitably leads to lower evaluable and – particularly – lower quality of the evaluation products and effects. This is illustrated clearly from the many evaluation reports, which indeed indicate that several aspects were not analysed thoroughly.

The score for the item whether ‘the key stakeholders are the party requesting an evaluation or at least are interested in it’ (3.1.2) is relatively good but still below the average score of the items in this component. Nevertheless in 85% of the interventions at least a majority of the stakeholders is interested or the requesting party. The preparedness towards and/or interest in an external survey can be explained in part by the strong result-oriented attitude of certain organisations (such as BTC) or countries (such as Rwanda), but also by the open and democratic organisation culture (particularly noticeable at NGAs). On the other hand, a negative attitude does not automatically imply a dislike of external evaluations. Interventions assuming that attitude often do it from the assumption that external evaluations do not have added value (with respect to internal evaluations and monitoring results); a negative experience with past external evaluations also played a part in a few cases.

The high scores for the compatibility of the expectations of the key stakeholders as regards the evaluation process and results with the reality value of those expectations (3.1.3 and 3.1.4), should be nuanced in the sense that their positive effect on the evaluable is not as impactful as we might assume. First of all, it is noted that the expectations of the key stakeholders are realistic because the level of ambition with respect to external evaluations usually is not very high. For instance, few interventions seek to actually invest in a thorough evaluation which focuses on impact or sustainability. On the other hand, if much is asked from the evaluators (as compared to the available resources), this implicitly means the acceptance of a trade-off in terms of

\(^{46}\) However, further down in the report we will find that results can differ on a country by country basis; see part 4: Comparative Analysis, below.

\(^{47}\) The group of (potential) users of evaluations may differ from the group of key stakeholders. The latter consists of stakeholders with a direct interest in the project and a close involvement in its implementation. The ‘users’ of the evaluation comprise these key stakeholders but may also include other parties, such as donors, members of a network of which the project forms part, national planning services, etc.
thoroughness of the analysis. As regards the mutual compatibility of the expectations there are not many problems, because the definition of the most important users and their expectations is not done systematically, as indicated before. The result is that possible differences between expectations and interests remain underexposed or are covered up by formulating broad evaluation objectives and themes.

The involvement of the users in the evaluation process (3.1.5) shows a good score in general, but not very good. It is important to note that only in exceptional cases, external evaluations are monopolised by one single stakeholder. Past experience has taught us that in such cases very little is done with the evaluation results. However, adequate involvement of the users puts high pressure on the evaluation implementation and evaluability. For this reason it is important that the involvement (preparation, implementation, follow-up) be measured adequately, in other words to consider the feasibility of the evaluation so as to avoid putting too much of an extra burden on the process. This exercise is not yet carried out in many interventions (not in the least because the users and their expectations and the objectives of the evaluation are not defined explicitly): most interventions quite intuitively choose for a certain form of involvement in, particularly, the preparations and follow-up, without checking whether this is the best choice. The involvement of key stakeholders (3.1.6) shows a slightly better score than the involvement of key users, which can be explained from the consideration that it is more obvious to involve these stakeholders in the evaluation. An important finding in this context is that just over one-third of the interventions explicitly involves target groups in evaluations. Some good practices show that proper involvement of users and key stakeholders is easier to establish through a broader approach, which focuses on involvement in the long term, not only in the context of evaluations, but via a systemic approach which starts at the project preparation and is continued during the implementation (see box 7).

Box 7: A broad approach in evaluations promotes high involvement from the key stakeholders

The PARZS project (Projet d'Appui au Renforcement des Zones et Départements Sanitaires du Mono-Couffo et de l'Atacora-Donga) carried out by BRC in the health industry in Benin has strongly applied the principles of outcome mapping (OM) in their planning, monitoring and evaluation system. The M&E systems are based on the definition of attitude changes and progress markers formulated by the different stakeholders and organisations involved in the project. The PARZS M&E system provides several moments of auto-evaluation and checks, which together guarantee good follow-up. Peer evaluations and the OM system are adapted to the local context and take account of the (monitoring) capacities and expectations of the parties involved. In the course of the project implementation, this system gradually has taken shape and is being used by the individual involved parties who were assigned responsibility. Various principles and elements of the system have also been incorporated into the M&E system of the health sector which was set-up on the basis of RBF (Results Based Financing).

In most interventions the relations between the key stakeholders are healthy (3.1.7). Where this is not the case, it often has to do with incidents that occurred in the past or with long-standing differences between stakeholders that sometimes go beyond the specific context. In most interventions the relations between the stakeholders are healthily critical, which means that there is no excessive friendliness. However, it is possible that a more 'specific' definition of the interests of key stakeholders and evaluation purposes and contents could increase critical behaviour between parties, although this need not necessarily be negative.

The fact that relations between key stakeholders in general are healthy is indeed an explanation for the high score (highest in this component) for the item 'it is possible to contact all of the key stakeholders without running a risk of mutual influencing’ (3.1.8). Furthermore it is also clear that in most interventions, the role of (external) evaluation in the development process and the position that should be taken towards the process, are
well understood. This also shows from the high score of the item which assessed the attitude of the key stakeholders as regards independent evaluation (3.1.9). In more than 80% of the interventions, almost all of the stakeholders have a positive attitude towards independent evaluations, something we feel is important from the perspective of evaluability.

### 3.3.2 The broader context

This last component of our analytical framework examines the broader context, i.e. the context beyond the direct environment of the interventions. It is clear that this context can affect the evaluability both in a positive and negative sense.

The broader context scores best of all of the components (the three items in this component rank among the 20% best scoring items (see also table A3 2 of Appendix 7)), despite the fact that the study included countries where the institutional and political environment can pose certain challenges for independent evaluations (table 3). The individual criteria show little differences. As could be expected, impact scores somewhat lower. This is also the case for relevance, which may have to do with the fact that when evaluating relevance, problems in the broader context have more weight than is the case for other criteria, which tend to concentrate more on the intervention in itself.

<table>
<thead>
<tr>
<th>Table 11: Overview of the findings on the broader context</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3.2 The broader context</strong></td>
</tr>
<tr>
<td><strong>Relevance</strong></td>
</tr>
<tr>
<td>4,10</td>
</tr>
<tr>
<td><strong>3.2.1 The broader institutional and political context is positive towards the independent evaluation</strong></td>
</tr>
<tr>
<td>4,40</td>
</tr>
<tr>
<td><strong>3.2.2 The social-cultural context at the level of the target groups allows for correct collection of information</strong></td>
</tr>
<tr>
<td>4,35</td>
</tr>
<tr>
<td><strong>3.2.3 Local expertise with the right profile for the evaluation is available</strong></td>
</tr>
</tbody>
</table>

In any case, the high scores are yet another illustration of the fact that independent evaluations generally can be carried out without any significant problems. However, conversations with the people involved have shown that problems will arise in almost every evaluation, but also that solutions are found quite easily and that these warrant a good implementation of the evaluation, hence also good evaluability.

This finding should nevertheless be put in perspective. First of all, we cannot exclude the fact that context factors which potentially put pressure on the evaluability have also affected our study, for instance through socially desired answers leading to positive scores. More in general the researchers have had too little time to thoroughly analyse the influence of the context, (for instance by having good conversations with a broad range of (potential) users of the evaluation). In fact an in-depth study at country level would be required to get good understanding of the influence of the context. As a result of the limited time and resources, the focus of this study on evaluation context had to be mainly on the technical rather than the political aspects of evaluation. Given such a ‘narrow’ interpretation of context it may not come as a surprise that solutions for any arising contextual problems are found relatively easily.

Furthermore we think that more in-depth reflection is required on the question whether or not the assumptions on carrying out evaluations independently have been too optimistic. In this context it is important, first and foremost, to review a few findings in relation to the previous component. For instance, we found (see 2.1) that defining the future users of the evaluation and their expectations and interests yields quite a low score. Since this aspect receives too little attention, there is quite a big a risk that the
set-up and development of evaluation exercises occurs from ‘within the system’. After all, as is also the case for independent evaluations, such independence relates mainly to the implementation – by independent experts – of the evaluation, whereas those experts hardly (or not at all) address the definition of the evaluation objectives and questions that the evaluation should provide answers for in the first place, one of the reasons being that they are not involved in specifying the reference terms. Similarly, it is also possible that ‘independent’ evaluations get to be controlled to a far-reaching extent and particularly avoid or underexpose controversial subjects (which might embarrass certain parties, for instance). If evaluations fail to deal with delicate subjects, then clearly there will be few or fewer problems in the area of institutional ad political context. We can speak of a type of ‘self-censorship’ here that also widely occurs in other situations. On the other hand, competent and smart evaluators can try to also pay attention to out-of-the-box aspects. In any case it is clear that such context puts pressure particularly on the learning potential but also on the general relevance of an evaluation. This finding invites us to add a nuance to the definition of evaluability, in the sense that evaluability also implies that everything that must be evaluated in principle, in fact can be evaluated.

The score for the item ‘The social-cultural context at the level of the target groups allows for correct data collection’ (3.2.2) also yields a high score. This does not mean that there are no challenges at target group level in this context, but it does mean that they are (sufficiently?) recognised and that one has (or thinks that one has) solutions for these challenges. In addition, a lot of (evaluation) experience has been gathered in recent years in relation to specific ‘difficult’ themes, such as gender-related violence, for instance in East-Congo, which makes that the problems are smaller than a few years ago. However, a risk identical to the one described in the previous paragraph applies in this area, namely that certain delicate subjects are being avoided or not actually dealt with. Furthermore, the research team is under the impression that the methodological challenges of research at a target group level are often underestimated. After all: the fact that one keeps good relations with target groups is no reason for a less strict application of the methodological and other requirements for proper research. In addition, interventions are becoming ever more complex, with a multitude of stakeholders, and for this reason they require ample evaluation expertise. And finally, evaluation or research fatigue occurs in many cases, with the potential of delivering highly misleading results.

Only in exceptional cases do interventions experience difficulty in contracting evaluators with the required profile (sufficient technical and methodological expertise; independence; availability at market rates) (3.2.3). The positive score is definitely an illustration of the increasing availability of particularly local expertise, which is important from the evaluability perspective. However, this requires nuancing as well. The fact that the 40 interventions in our sample relate mainly to ‘traditional’ interventions, implies that finding good experts will not be problematic. Experience with the evaluation of less traditional types of cooperation, such as budget aid, shows that this is not always the case. Another explanation could be that the (frequent) absence of firm objectives and evaluation questions implies that the requirements as regards the necessary expertise are not very firm either.

### 3.3.3 Practical elements

For the sake of completeness we wish to restate that there is a third component of importance in the analysis of the evaluation context and the determination of the evaluability. It involves a few practical elements with a mostly direct effect on evaluability, such as safety situation, weather circumstances, the state of the local infrastructure, the geographic size of the intervention area and the access, etc. Timing of the evaluation can also affect the evaluability quite strongly: field work in evaluations

---

48 We have not been able to study this important aspect sufficiently to be able to make far-reaching statements on this issue.

49 As early as the specifications phase we decided not to include budget aid in this study.
preferably should not coincide with elections, religious events, holiday periods, hectic periods in the intervention, etc. Furthermore, certain events (conflicts in organisations or interventions) or similar initiatives (e.g. from other donors) can cause complications. In view of the highly intervention-specific character of these elements we chose not to include them in our analysis.
4 Comparative analysis

Dit hoofdstuk bouwt verder op de analyse in hoofdstuk 3 uitgevoerd en probeert na te gaan in hoeverre er verschillen zijn in de evalueerbaarheid op basis van bepaalde parameters die reeds eerder bij de bepaling van de steekproef werden gehanteerd. Meer bepaald gaat het daarbij om (1) de vier landen, elk met 10 interventies vertegenwoordigd in de steekproef, (2) de graad van complexiteit van de interventies, waarbij we een onderscheid hebben gemaakt tussen interventies met een ‘complexe’ en ‘minder complexe’ TOC, waarbij 40% van de interventies als ‘minder complex’ werden gecatalogeerd en 60% als ‘complex’, en (3) het type actor, waarbij 25% van de interventies onderdeel uitmaakt van de bilaterale samenwerking, 50% van de samenwerking via NGOs en vakbonden, en 25% via andere actoren (APEFE/VVOB, universitaire samenwerking, BIO, ITG, BOS+, VVSG). Deze analyse steunt in grote mate op een statistische verwerking van de scores die in bijlage 7 verder wordt toegelicht.

4.1 Comparison of evaluability at a country level

The following table presents the scores at a component level and for the three large dimensions of the study framework. It is important to note that similar scores at a component level may disguise differences between countries at an item level. These potential differences are analysed by comparing the average scores per DAC criterion at component and item level between the four countries.

The table below shows that there is no clear general pattern as regards the differences between the evaluability scores of the four countries. There is a clear difference where it would be most expected, namely under the third dimension which analysed the effect of the context on the evaluability: DRC and subsequently, Rwanda score less than the other two countries. The relatively small differences in the scores for the other dimensions and components suggest that other parameters than the country could be of greater importance for explaining the differences between the scores (see discussion under 4.2 and 4.3 below).

50 Voor meer informatie i.v.m. de steekproefbepaling, zie hoofdstuk 2.3.1, bijlage 4 (lijst van alle opgenomen interventies) en bijlage 2 waar meer informatie wordt gegeven m.b.t. de onderzoeksmethodologie.

51 Initieel wilde de studie een vergelijking maken tussen interventies in “harde” en “zachte” sectoren vanuit de veronderstelling dat interventies in harde sectoren wellicht meer evalueerbaar zijn dan in zachte, maar dit bleek niet werkbare. Na een vrij lange interne reflectie bleek vooral de complexiteit van de TOC van de interventie belangrijk en werd overgegaan tot een onderscheid tussen ‘complexe’ en ‘minder complexe’ interventies, met name interventies met een complexe en minder complexe TOC. Zie punt 2.3.1 en punt 2 van bijlage 2 (Beschrijving van de aanpak en onderzoeksmethodologie) voor meer details.

52 This part is based, among other things, on the four country notes that were prepared and shared with the organisations involved. However, the notes are not official.
Table 12: Overview of the evaluability scores per country

<table>
<thead>
<tr>
<th>Dimension 1 (intervention plan)</th>
<th>Belgium</th>
<th>Benin</th>
<th>DRC</th>
<th>Rwanda</th>
<th>Evaluability index (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 The underlying analysis</td>
<td>3,14</td>
<td>2,94</td>
<td>3,24</td>
<td>3,32</td>
<td>3,16</td>
</tr>
<tr>
<td>1.2 The intervention logic and theory of change</td>
<td>2,70</td>
<td>2,63</td>
<td>3,58</td>
<td>2,95</td>
<td>2,96</td>
</tr>
<tr>
<td>1.3 The proposed M&amp;E system</td>
<td>2,55</td>
<td>2,18</td>
<td>2,34</td>
<td>2,98</td>
<td>2,51</td>
</tr>
<tr>
<td>1.4 Consistency and adaptation of the intervention logic and theory of change</td>
<td>4,08</td>
<td>3,17</td>
<td>3,90</td>
<td>3,03</td>
<td>3,54</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension 2 (implementation practice)</th>
<th>Belgium</th>
<th>Benin</th>
<th>DRC</th>
<th>Rwanda</th>
<th>Evaluability index (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Basic information as regards the intervention implementation</td>
<td>2,88</td>
<td>2,68</td>
<td>2,90</td>
<td>3,08</td>
<td>2,88</td>
</tr>
<tr>
<td>2.2 M&amp;E system in practice</td>
<td>3,08</td>
<td>2,66</td>
<td>3,20</td>
<td>3,22</td>
<td>3,04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension 3 (context)</th>
<th>Belgium</th>
<th>Benin</th>
<th>DRC</th>
<th>Rwanda</th>
<th>Evaluability index (°°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Attitude of the key stakeholders</td>
<td>4,17</td>
<td>4,18</td>
<td>3,85</td>
<td>3,45</td>
<td>3,91</td>
</tr>
<tr>
<td>3.2 Broader context</td>
<td>4,56</td>
<td>4,60</td>
<td>4,20</td>
<td>3,32</td>
<td>4,17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>General score for evaluability (°°)</th>
<th>Belgium</th>
<th>Benin</th>
<th>DRC</th>
<th>Rwanda</th>
<th>Evaluability index (°°)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3,28</td>
<td>3,08</td>
<td>3,29</td>
<td>3,28</td>
<td>3,23</td>
</tr>
</tbody>
</table>

(°) The score for the general evaluability index is calculated on the basis of indexes (not averages). The scores calculated for the countries are averages (not indexes). Consequently, it is possible that the general evaluability score deviates slightly from the average for the four country scores. The differences are very tiny and do not change the contents of this analysis.

(°°) For the general evaluability score a weighted average was calculated in which dimensions 1 and 2 are weighed equally and dimension 3 counts for half of each of the other dimensions.

The more in-depth statistical analysis at a component level and particularly at item level shows a few significant differences between countries (which will be discussed below), but an unambiguous pattern cannot be distinguished (see Table A2 3 in Appendix 7 for more details). A closer study of the above table, a comparison of the average scores achieved per DAC criterion at component and item level\(^54\), combined with the results of the statistical analysis, allow us to better distinguish the differences between the countries.

The results of the further analysis are in line with the observations for table 12; the clearest difference can be found at the level of the broader context (3.2), where the lower score for the interventions in Rwanda stands out for all five criteria\(^55\). At an item level we find only statistically significant differences in relation to the institutional and political context (3.2.1)\(^56\) The scores indicate that this context affects the evaluability of the interventions in Rwanda in a negative way because the implementation of independent evaluations can be seriously hampered. However, the social-cultural context at the level of the target groups (3.2.2) and the availability of local expertise (3.2.3) are not factors that, given a comparison between countries, have a different effect on the evaluability of the interventions. The relatively good scores for DRC also stand out. These scores do not differ significantly from the scores for Benin and Belgium; in other words: despite the political and institutional troubles in the country, this study finds no

\(^{54}\) The average scores per country can be found in table A2 1 of Appendix 7. The Appendix also contains the results and a brief description of the related statistical tests.

\(^{55}\) ...nevertheless, the Rwanda score also remains high in general (especially compared to the other dimensions). It is important to point towards the observations made in the previous chapter about the way in which 'context' was approached in the study (see the discussion in part 3).

\(^{56}\) See also Appendix 7 for more information about the statistical analyses that were carried out.
indications that those contextual factors would have a (more) negative effect on the evaluable of interventions in the country. For the other significant differences it is more problematic to find and unambiguous interpretation since those differences do not always occur between the same countries. What is remarkable, however, is that when significant differences occur, Rwanda often is one of the best scoring countries. Consequently it seems that in Rwanda the context affects the evaluable of interventions in both a positive (1.1, 1.2, 1.2, 2.1 and 2.2) and a negative (3.2 and 1.4) way. This can be explained from the very clear policy context (both on paper and in practice), which sets Rwanda apart from the other countries. Among other things the country's performance culture ensures: that a clear rationale and problem situation can/must be formulated for each intervention; that the relation between analysis and intervention objectives is/must be clear and unambiguous; and that, more than elsewhere, the interventions (must) have an eye for the relation between intervention and the country's policy. This extensive performance culture also ensures that M&E systems, particularly when they are directed towards upward accountability, are better developed. The need for synchronisation with local systems also affects the sustainability of interventions (see the high scores for Rwanda for sustainability criterion under components 1.2 and 1.3).

As indicated earlier, given the limited number of observations for this component we must be careful in interpreting the component related to the consistence and adaptation of the intervention logic and the theory of change (1.4). As regards the evaluable of efficiency, impact and sustainability, we did not find any significant differences between the countries. We did not find a significant difference between indicating and substantiating any changes (1.4.1) either. The scores differ only for the relevance and effectiveness criteria for the items related to incorporating the changes in the M&E system (1.4.3) and the availability of information about the vision and opinions of the key stakeholders as regards these changes (1.4.2). The interventions in Rwanda in particular show a lower score for these items. An extensive performance culture may have affected the scores for answering this component in several ways. For instance, the strong stress on ‘quick wins’ could cause less openness in interventions to report any problems, failures and related changes in interventions. This is closely related to the aforementioned tendency of self-censorship. In addition, given the speed with which (government developed) reforms are implemented effectively in Rwanda, it is impossible for interventions to report all the changes this brings about. It does, however, cause a gap between what exists on paper and what is real, which has a negative effect on evaluable.

We don't have an easy explanation for the relatively high score in DRC with respect to the quality of intervention logic and theory of change. A possible explanation could be that 7 out of 10 interventions at DRC constitute the second or third phase of earlier interventions, which could have contributed to gradual improvement of the intervention logic and theory of change.

57 It is important to note that the differences between the countries can also relate to differences in the area of the intervention channel (bilateral v. other) and the sector/type of intervention (services v. lobby and advocacy for instance). The study attempted to create a similar sample per country, but succeeded only partly as a result of the limited number of interventions per country and the multitude of parameters involved; furthermore, the sample is too small to distinguish any interaction effects. In this context the situation in Belgium is quite special. Many interventions in Belgium differ greatly from those in the South because they start by developing an offer for their target groups, who are ‘consumers’ rather than actively involved stakeholders; consequently, they are less accessible and usually less interested in evaluations. Other specific problems in some of these interventions include the changing target groups and the difficulty of tracing target groups at a later stage who were reached at earlier stages.

58 This also applies to comparative analyses according to other parameters that follow below, and will not be repeated there.
4.2 Comparison of evaluability based on the complexity of the interventions

The following table presents the scores at a component level and for the three large dimensions of the study framework. As was the case for the analysis at country level, it is important to note that similar scores at component level can disguise differences at item level. These potential differences are analysed by comparing the average scores per DAC criterion at component and item level between complex and less complex interventions.

Table 13: Overview of the evaluability scores for interventions with a complex and less complex TOC

<table>
<thead>
<tr>
<th>Item</th>
<th>Less complex interventions</th>
<th>Complex interventions</th>
<th>Evaluability index (*)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dimension 1 (intervention plan)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1 The underlying analysis</td>
<td>2,87</td>
<td>3,30</td>
<td>3,16</td>
</tr>
<tr>
<td>1.2 The intervention logic and theory of change</td>
<td>3,03</td>
<td>2,92</td>
<td>2,96</td>
</tr>
<tr>
<td>1.3 The proposed M&amp;E system</td>
<td>2,11</td>
<td>2,78</td>
<td>2,51</td>
</tr>
<tr>
<td>1.4 Consistency and adaptation of the intervention logic and change</td>
<td>3,03</td>
<td>3,67</td>
<td>3,54</td>
</tr>
<tr>
<td><strong>Dimension 2 (implementation practice)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1 Basic information as regards the intervention implementation</td>
<td>2,68</td>
<td>3,03</td>
<td>2,88</td>
</tr>
<tr>
<td>2.2 M&amp;E system in practice</td>
<td>2,95</td>
<td>3,10</td>
<td>3,04</td>
</tr>
<tr>
<td><strong>Dimension 3 (context)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1 Attitude of the key stakeholders</td>
<td>3,44</td>
<td>3,80</td>
<td>3,66</td>
</tr>
<tr>
<td>3.2 Broader context</td>
<td>4,20</td>
<td>4,15</td>
<td>4,17</td>
</tr>
<tr>
<td><strong>General score for evaluability (°°)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3,04</td>
<td>3,34</td>
<td>3,23</td>
</tr>
</tbody>
</table>

(*)The score for the general evaluability index is calculated on the basis of indexes (not averages). The scores calculated for complex and less complex interventions are averages (not indexes). Consequently, it is possible that the general evaluability score deviates slightly from the weighted average of the scores of less complex (40%) and complex (60%) interventions. The differences are very tiny and do not change the contents of this analysis.

(°°) For the general evaluability score a weighted average was calculated in which dimensions 1 and 2 are weighed equally and dimension 3 counts for half of each of the other dimensions.

The table above shows that 'complex interventions’ (i.e. interventions with a complex TOC) generally (namely for 6 of the 8 components) score a little better than less complex interventions. This is a surprising conclusion because we initially started from the assumption that complex interventions are harder to evaluate than less complex ones. As a matter of fact this hypothesis seems to be true only for component 1.2, where we established that the score for 'the intervention logic and the theory of change' is somewhat higher for less complex interventions, but the difference is so minimal that in fact it does not allow us to draw any conclusions. If we can at all give an explanation for the finding at the level of this component, then it would be that it could have to do with the fact that in complex interventions it is less evident to identify critical elements because there is less empirical evidence as a basis for an assumption.

A general explanation for the relatively good score of complex interventions could be that exactly because of the TOC complexity, in these interventions one will pay more attention to, for instance, the underlying analysis and the M&E system. This would imply that the stakeholders pay more attention to ‘learning’ because they realise that little empirical evidence is available as yet and analysis and the M&E system therefore require more attention in order to allow for fine-tuning. On the other hand, from an accountability standpoint there is an awareness that the donor may be more sceptical towards ‘difficult’ complex interventions whose effectiveness is harder to prove.
To get a better understanding of the differences as regards evaluability between interventions with a complex TOC and those with a less complex TOC, we made a comparison between the average scores achieved per DAC criterion at component and item levels\(^59\). In the comparison between interventions with a complex and less complex TOC, we found significantly different scores only for 2 components: the proposed M&E system (1.3) and the attitude of key stakeholders towards independent evaluations (3.1). The differences are at the level of relevance, effectiveness and efficiency for the first item, and only at the level of effectiveness and efficiency for the second.

Further analysis of these two components shows that interventions with a complex TOC score significantly higher. When we extend the comparison to the underlying items, we see that the trend is maintained, with the exception of the item about the identification of the users of an evaluation and their expectations (3.1.1). For interventions with a complex TOC, identifying the users of an evaluation and their expectations may be harder, for instance because of the large number of different stakeholders with different expectations in such interventions; because of the fact that there is often less experience with a (more) complex theory of change, as a result of which there can be more uncertain factors and identification of the users and their expectations is harder, etc.

It seems hard to explain why interventions with a less complex TOC for the indicated DAC criteria score significantly lower for the proposed M&E system (1.3) and lower for the item concerning the translation of the underlying intervention logic into the proposed M&E system... Maybe one of the reasons is that such interventions are carried out mainly by less experienced organisations, as a result of which technical components or items also yield lower scores\(^60\).

As already indicated in part 3.1 of the previous chapter, the higher score for interventions with a more complex TOC as regards the underlying compatibility of the expectations (3.1.3) can in part be explained as follows: there are few or fewer problems because, as shown by the score for 3.1.1, defining the main users and their expectations is not done systematically. As a result, any differences in expectations and interests of these users remain underexposed.

The higher score for interventions with a complex TOC with respect to the item whether the key stakeholders are the party requesting or at least showing an interest in an evaluation (3.1.2) could have various causes (as already indicated in part 3.1 of the previous chapter): interventions with lower scores for this item do not have an explicitly negative attitude towards external evaluations because the lower score could indicate the assumption that external evaluations will not offer any added value compared to existing internal evaluations and the available monitoring results. If, for instance, it is easier for interventions with a less complex TOC to monitor results, then it is possible that in these interventions the added value of independent evaluations is rated lower.

In most interventions the relations between the key stakeholders are healthy. It could be that interventions with a less complex TOC score lower on this item (3.1.7) because, as already discussed above, the interventions with a complex TOC have created a less explicit definition of the users and their interests and expectations with respect to evaluation objectives and contents. As indicated under point 3.1 of the previous chapter, it is possible that a more ‘conscious’ definition of interests and expectations as regards evaluation objectives and contents would heighten tensions between stakeholders. On the other hand, if we combine the result of ‘the relations between the key stakeholders are healthy’ with the high score (highest in this component) for the item ‘it is possible to contact all of the key stakeholders without running the risk of mutual influencing’, it could also mean that given the uncertainty, interventions with a more complex TOC are more interested in different opinions (for the purpose of learning from them).

---

\(^{59}\) The average scores can be found in table A2 7 of Appendix 7. Appendix 7 also contains a brief description of the related statistical tests.

\(^{60}\) However, the sample if too small to make any reliable assumptions on this issue.
At the level of the evaluability of impact and sustainability there are no significant differences between interventions with a complex TOC and those with a less complex TOC. This could be explained from what has been written about those levels in the above analyses, namely less specification of these levels compared to the other three DAC criteria. We would like to add that as regards the components comprising the underlying analysis (1.1) and the underlying intervention logic and the theory of change (1.2), we found no significant differences between interventions with a complex TOC and interventions with less complex TOC for any of the relevant DAC criteria. Although interventions with a complex TOC would sooner be associated with more problems as regards, for instance, the development of the intervention logic and TOC, in practice this does not seem to lead to analyses and/or intervention logic which affect the evaluability of these interventions in a negative way. All interventions show relatively good scores for these components. Field visits also showed that despite the problems, interventions with a more complex TOC make a big effort to specify underlying analyses and theories of change and to translate them into consistent intervention logic.

4.3 Comparison of evaluability at stakeholder level

The following table presents the scores at a component level for the three large dimensions of the study framework. As was the case for the analysis at country and complexity level, it is important to note that similar scores at component level can disguise differences at item level. These potential differences are analysed by comparing the average scores per DAC criterion at component and item level for the three types of stakeholders.

Table 14: Overview of the evaluability scores per stakeholder

<table>
<thead>
<tr>
<th>Item</th>
<th>BTC</th>
<th>NGOs and trade unions</th>
<th>Other</th>
<th>Evaluability index (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dimension 1 (intervention plan)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1 The underlying analysis</td>
<td>3,52</td>
<td>3,20</td>
<td>2,59</td>
<td>3,16</td>
</tr>
<tr>
<td>1.2 The intervention logic and theory of change</td>
<td>4,10</td>
<td>3,75</td>
<td>2,90</td>
<td>3,65</td>
</tr>
<tr>
<td>1.3 The proposed M&amp;E system</td>
<td>3,00</td>
<td>2,98</td>
<td>2,90</td>
<td>2,96</td>
</tr>
<tr>
<td><strong>Dimension 2 (implementation practice)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1 Basic information as regards the intervention implementation</td>
<td>2,86</td>
<td>3,03</td>
<td>2,62</td>
<td>2,88</td>
</tr>
<tr>
<td>2.2 M&amp;E system in practice</td>
<td>2,95</td>
<td>3,10</td>
<td>2,78</td>
<td>3,04</td>
</tr>
<tr>
<td><strong>Dimension 3 (context)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1 Attitude of the key stakeholders</td>
<td>3,74</td>
<td>3,72</td>
<td>3,44</td>
<td>3,66</td>
</tr>
<tr>
<td>3.2 Broader context</td>
<td>4,40</td>
<td>4,14</td>
<td>4,00</td>
<td>4,17</td>
</tr>
<tr>
<td><strong>General score for evaluability (°°)</strong></td>
<td>3,41</td>
<td>3,30</td>
<td>2,86</td>
<td>3,23</td>
</tr>
</tbody>
</table>

(°)(°) The score for the general evaluability index is a score calculated on the basis of indexes (not averages). The scores calculated for the stakeholders are averages (not indexes). Thus, it is possible that the general evaluability score deviates slightly from the weighted average of the interventions; interventions of 'BTC', of 'NGOs and trade unions' and of 'Other' constitute 25%, 50% and 25%, respectively, of the sample. (°°) For the general evaluability score a weighted average was calculated whereby dimensions 1 and 2 were weighed equally and dimension 3 counts for half of each of the other dimensions.

The above table shows that compared to the two previous parameters, the different stakeholders or ‘channels’ yield bigger differences in terms of evaluability. Obviously they do not occur as much in the field of the third dimension (effect of the context), as in the two other dimensions and particularly in dimension 1 (the intervention plan). The data in the table shows that the evaluability of BTC and NGO/trade unions interventions is much higher than for other stakeholders, which does not imply that these differences should also apply to individual interventions within each group of stakeholders; this
applies *a fortiori* for the group ‘Other stakeholders’ which comprises a broad range of organisations and practices.

A closer study of the above table and a comparison of the average scores achieved per DAC criterion at component and item level\(^{61}\) allow us to better distinguish the differences between stakeholders. The results of the statistical tests (see Appendix 7) show that the three stakeholders yield significant differences only at the level of the subject of the intervention (part 1). On the other hand, the context and practice of the implementation of the intervention of these three intervention channels have no significantly different effect on the evaluability of the interventions.

Furthermore, the statistical analysis shows that the differences concerning the subject of the intervention are situated at the level of the underlying analysis (1.1) and the proposed M&E system (1.3).

The difference as regards the underlying analysis (1.1) manifests itself mainly in differences in items concerning the definition and description of target groups (1.1.1), the description of the role of the target groups and of the stakeholders (1.1.3 and 1.1.4) and the presence or absence of a proper gender analysis (1.1.5). As regards the first three items, the high scores of BTC interventions and the low scores of the interventions in the ‘Other’ intervention channel catch the eye. Possible explanations include the fact that BTC demands that all interventions use the same, properly specified and detailed formats for writing intervention propositions and the general technical and financial file. These formats also pay sufficient attention to describing the target group and the role to be fulfilled by target groups and/or stakeholders in the intervention. Furthermore, both BTC (via MoreResults) and NGOs (via initiatives of the federations and change projects) pay plenty of attention to improving project management. This is in contrast with the intervention channel ‘Other’ which, in some cases, does not even require a project proposition and/or underlying analysis (for the purpose of applying for funds) and which uses less complete and more diverse formats. In addition, for some of these ‘Other’ stakeholders (e.g. cities and municipalities) development cooperation is only a side activity in which they do not specialise.

As regards the gender analysis (1.1.5) a higher score can be found for the intervention channel ‘NGOs and trade unions’. A possible explanation could be that these stakeholders traditionally focus more on reaching socially weaker target groups for which it is easier to link gender relations to underlying explanatory mechanisms, as a result of which the need for integration of gender analysis and translation into the M&E system becomes more apparent. The relatively weak score of the BTC interventions could be explained from the fact that during the planning and implementation stages of these interventions, gender but sporadically received attention. The attention for gender in the MoreResults manual does not differ significantly from that in earlier guidelines.

Furthermore, the three intervention channels did not provide significantly different scores for the items concerning the rationale of the intervention (1.1.2), the relation between the underlying analysis and the intervention objectives (1.1.6) and the position of the intervention with respect to the local sector policy (1.1.7). This is an additional indication that justifies the statement that good formats play an important role in the implementation of the plan, but that the interventions of the three different intervention channels do indeed have a rationale in which logical intervention objectives get preliminary attention and which takes into account the policy context of the intervention. No significant differences were found between the three intervention channels with respect to intervention logic and the theory of change (1.2). This finding indicates that the intervention channels show no differences in technical competence to incorporate interventions in a theory of change and to develop an intervention logic.

\(^{61}\) The average scores can be found in table A2.4 of Appendix 7. The Appendix also contains the results and a brief description of the related statistical tests.
As regards the differences found for the three intervention channels at the level of proposed M&E system (1.3), we can distinguish three different ‘trends’. First of all the interventions of the ‘Other’ intervention channel show a significantly lower score for the proposed M&E system, which also returns in the items concerning operationalisation of the intervention results (1.3.1), the work method for monitoring and evaluating the intervention results (1.3.4), the description of the financial and HR resources of the M&E system (1.3.7), the presence or absence or use of an MIS (1.3.8) and the M&E system as a consistent translation of the intervention logic (1.3.3). As mentioned earlier, different requirements as to reporting and report format contents apply to this type of intervention channels. Within this type of intervention channels, it is not always necessary to define operationalised objectives, nor to monitor them in a systematic way. In this context, it is not surprising either that practical issues such as the resources available to the M&E system are less exposed, since this type of intervention channels seldom have an existing M&E culture and practice. This general finding obviously does not preclude that certain individual interventions may have proposed and developed good M&E practices. In this context it should be noted that the scores for the M&E system in practice are much closer to each other than for the proposed M&E system.

Secondly, we note that the interventions in the ‘NGO and trade unions’ intervention channel score significantly higher for the clear description of the working method to monitor assumptions (1.3.5). This could be explained from the fact that for the interventions of NGOs and trade unions, the government policy belongs to the external context and that these interventions are aware of the fact that the government policy is part of the external assumptions that may affect the interventions. On the other hand, the government policy for the bilateral interventions belongs much less to the external assumptions (more to the internal risks) and furthermore, other external assumptions with a potential effect on the interventions are much harder to identify. However, interventions in the ‘Other’ intervention channel are less tied to a government policy and could find it harder to identify other external assumptions.

Thirdly, the interventions in the ‘BTC’ intervention channel yield a higher score for a clear description of how the M&E system is fine-tuned to the local/national M&E systems (1.3.9). This difference is easily explained from the implementation methods used for BTC interventions and in which fine-tuning to local/national politics and the associated monitoring methods forms an explicit point of attention.
5 Conclusions and recommendations

5.1 Main conclusions

The following conclusions and lessons learnt are presented in two parts. The first part mainly constitutes a summary of the main findings. The second part is more analytical in nature.

5.1.1 Synthesis of the main results and findings

Meaning and importance of ‘evaluability’ in the development practice

Evaluability is defined as the extent to which an activity or a programme can be evaluated in a reliable and credible fashion. In that context, the definition of the 'feasibility' of an evaluation is extended to checking the desirability of an evaluation. This means that evaluability mainly concentrates on three dimensions: (the quality of) the intervention design; (the quality of) the intervention implementation (including the M&E system); and the role of the broader intervention environment.

The first and more evident relevance of evaluability starts from the consideration that performing an evaluability test (or assessment) requires but a fraction of the resources needed for carrying out the actual evaluation, whereas its potential added value is significant. In other words, an assessment of the evaluability will show whether an evaluation is desirable and feasible (at a certain time; in a certain context) and if yes, under which conditions.

However, adequate use of ‘evaluability’ and the performance of an evaluability assessment should not necessarily be associated with the set-up of evaluations – they have a broader usefulness which could benefit the management of development interventions as a whole. For instance: an analysis of the evaluability may improve the intervention design or make an important contribution to the design and the development of an M&E system and provide important input as regards desirability, timing, approach and objectives of an evaluation and thus contribute to the usability and effective use of the evaluation result, hence of the actual usefulness of the evaluation.

The above shows the importance of the concept evaluability as well as of the instrument ‘evaluability test’: their potential application goes beyond (organising) evaluations and concerns the entire intervention management. The study found that until now, the concept evaluability and the instrument evaluability assessment are little known and used in the Belgian development cooperation environment. Elements of evaluability tests are practiced here and there – without naming them as such – but there is no systematic use.

This finding is significant because ‘evaluation’ has become increasingly important in recent years and has become a fixed part of the management practice of interventions, programmes and the development cooperation as a whole: the relevance and usefulness of evaluations are not questioned and nowadays stakeholders simply cannot afford not to evaluate. In other words: evaluation has become an obligation. However positive this development may be, it also holds the risk that evaluations turn into ritual exercises without any true involvement of the key stakeholders. A conscious use of the concept of
evaluability and of the instrument evaluability test, linked with the option of arriving at well-founded statements about the desirability of an evaluation, may in this context prove to be an important instrument towards improved and more relevant implementation of the role and function of evaluation(s) in the environment of development cooperation.

A general view of evaluability

The forty analysed interventions returned an average score – based on the sixty-two items in the study framework – just above the centre point of the scale (see table 3). Although we cannot attach any absolute meaning to the scores, the general score forms a good indication of the main finding of these studies: interventions generally have a number of strong points, but also contain points that require additional work if they want to improve their evaluability. Another remarkable finding is that the (statistical) distribution of the scores approximates normal distribution, but is characterised by large spread: the lowest scoring interventions don't even reach the medium score of the highest scoring interventions. Because ‘evaluability’ is closely linked to the intervention management practice, this finding implies that the Belgian stakeholders and types of intervention show large differences in terms of management, despite control of the funding government.

Of the three studied dimensions, i.e. intervention plan, implementation practice and context, the latter scored considerably higher than the other two. This illustrates, again, the margin for improvement of the dimensions that stakeholders have most control of (intervention design and implementation practice). The score further suggests that, at least for the 4 countries in the study, the challenges at context level do not constitute a heavy burden for the evaluability. This interpretation requires nuancing, however, since we were unable to make an in-depth analysis of any restrictions at context level.

If we consider the general scores per evaluation criterion, the lower scores for sustainability and particularly impact jump to the fore: despite increased attention for sustainability it is still insufficiently integrated in the management systems; as far as impact is concerned, a combination of factors makes for difficult evaluability. On the other hand, the scores for effectiveness and particularly efficiency are considerably higher than the medium value of the developed scale, which forms an indication – across the different dimensions and components – of quality of the intervention management, in particular with respect to monitoring and evaluation.

The intervention design

The generally good score for the underlying analysis points towards a certain tradition and skill as regards the implementation of analyses, which appears to stem from internal processes as well as requirements from the donors for whom such analysis constitutes an important item in the assessment of grant applications. Furthermore, in many cases the initiators can rely on past experience (follow-up interventions). An important positive finding is the broad attention for policy context in many analyses, stemming in part from better networking and platform activities between the Belgian stakeholders. However, there are a few important bottlenecks in the field of analysis too: the limited attention for target groups, for gender and for (factors of) social differentiation in general; the limited attention for key stakeholders (other than target groups); and the narrowing of the analyses, which often serve to account for strategic and policy choices rather than forming a basis thereof. These shortcomings have a negative effect on evaluability because they interfere with the assessment of the relevance of the policy choices made and make it difficult to assess who will be reached effectively (as compared to the planning) and whether crowding out effects have occurred at target group level.

The intervention logic and theory of change component yields a relatively low score, which can be attributed to the limited attention for higher levels in the ends-means
chain. In practice, attention focuses on the implementation level (particularly in screening) and there seems to be little incentive hence interest in superseding that level: not only does the implementation level have a greater and direct usefulness for the organisations involved, the funding government is in fact more interested in accountability (as regards the correct use of the funds) and furthermore, the developed formats pay little attention to the impact level. An additional problem is that impacts, once formulated, are situated at great distance in the ends-means chain, and that the theory of change thus contains a missing middle. The existing intrinsic challenges, the absence of external incentives and the limited knowledge of and insight into existing (and realistic) methods and opportunities for impact evaluation cause important problems as early as the preparatory phase, which make the evaluability of impact and the actual implementation of impact evaluations (i.e. as regards the outcomes and direct effects of these outcomes) at a later stage problematic. On the other hand, the good focus on ‘implementation’ ensures a good score for the efficiency criterion which actually extends into the actual practice (e.g. in the good connection of the use of resources to the outputs).

The quality of the M&E proposition in the intervention proposition yields a weak score. There are various explanations for this: the absence of incentives to invest, ex ante, in the description of the M&E system; the fact that attention for M&E has increased only recently (many stakeholders have put a lot of work into developing an M&E policy) but has not reached the intervention level; the choice of some stakeholders to invest in developing an M&E system only at the start of an intervention. From an evaluability perspective the availability, ex ante, of an M&E design is important only insofar as this implies a good M&E practice later on. In practice, good ex ante attention does seem to yield good practice, however we also found interventions in which a weak ex ante score gets to be corrected later on by good practice. Furthermore, well-prepared (paper) M&E propositions also appear to be important in the light of the high level of staff turnover and the high number of task changes in many interventions. An finally, it is clear that initial bottlenecks (e.g. limited attention to impact and gender; no specific description and attribution of resources for monitoring and particularly evaluation) often resurface at a later stage. This applies – a fortiori – for the finding that an actual M&E system seldom exists; although in many cases components of such a system are proposed, a lot of work is still required to warrant appropriate internal coherence.

Changes in intervention logic and theory of change in the course of the implementation are usually signalled appropriately and justified – a positive finding from the perspective of evaluability. However, in many cases these changes are not transferred to the M&E system, which forms an indication of the weak institutional incorporation of the system.

The implementation practice

The basic information related to the progress of the intervention implementation is relatively readily available but, for reasons of language and sometimes of complexity, is not always accessible for (or known by) all of the key stakeholders. Furthermore this information reflects the information which is available for the intervention plan: it focuses mainly on the implementation level, remains at the surface with respect to the aspects of social differentiation (including gender) and pays little attention to higher levels in the end-means chain. In addition the information suggests an inward-looking focus in the intervention management, with little attention for external hypotheses, options for developing a counterfactual, or the approach and quality of the data collection process. The rather weak score for this item must be put in perspective, however, because in an important number of cases the field visits revealed good practices which could not be traced back to the basic documents.

The M&E system in practice is probably the component with the strongest effect on the evaluability on an intervention. After all: weaknesses in other areas can to some extent be compensated by a coherent and well-functioning M&E system, but conversely,
particularly the practical evaluability of an intervention will be hampered by an ill-functioning M&E system. In light of this consideration the fact that the actual M&E practice scores considerably higher than the proposed M&E system is a positive finding. On reflection, the M&E practice often appears to be a 'work in progress', limited to the operational level for reasons partly explained above and originating from imperfections in the intervention design and the good, but all in all relatively recent attention for M&E, which makes that policy decisions in this area have not yet been extended to the local level. Furthermore a number of mechanisms seem to be in play that affect the characteristics of the M&E practice: the close relation between operational monitoring and the main tasks of programming staff (contrary to result-oriented monitoring which is often more demanding and 'more at a distance'); the lack of indicators at impact level and sometimes at outcome level, resulting in underexposure of these levels in the M&E practice; the absence of (attention for) specific competence development in the area of M&E; the fact that for many stakeholders, M&E is an added responsibility; the lack of time, resources and procedures to put M&E results to good use for thorough analysis, decision-making and adjustments. However, these weaknesses are compensated in part by an organisation culture which is oriented towards reflection and learning. Furthermore, it seems that until now, M&E policy and practice focus mainly on ‘monitoring’ at the expense of evaluation. This has several reasons: the fact that the M&E policy is not yet manifest in practice and that therefore, logically, attention focuses on monitoring at first; the greater methodological requirements for evaluation in comparison to monitoring; and the doubt (particularly in organisations with an appropriate monitoring system and in complex interventions) as regards the possible added value of external evaluations, as a result of which such evaluations are not supported with the necessary resources (see also below).

A final but important finding is related to the strong inward-looking attitude in respect of M&E practice among many of the Belgian development stakeholders and which translates mainly in a limited role in M&E policy and implementation for the intervention stakeholders (other than the responsible organisation and the intervention team); the reduction of accountability to donor accountability; and the limited attention for fine-tuning M&E to local and national systems (including those of the partner) and for the possible role that Belgian stakeholders can or must take in setting up or improving these systems.

Summarising, it appears that the current implementation practice holds opportunities as well as limitations in terms of evaluability, in which context it is important to realise that this study provides but a snapshot of a practice which is subject to considerable change.

The influence of the evaluation context

The attitude of key stakeholders as regards (independent) evaluations is an important factor of evaluability. The study revealed that this attitude is a direct obstacle in but a few cases. Key stakeholders are generally positive about and interested in external evaluations and also know what type of role is expected from them (e.g. respect for the autonomy of the evaluator; no influencing). However, reality is a little more complex than meets the eye: in practice it appears that a very limited number of stakeholders gets to be involved in the evaluation process and that the interests and expectations of a wide range of stakeholders is not included in the evaluations (e.g. as a result of an unclear M&E policy, a weak evaluation practice and scepticism with respect to the added value of external evaluations). The result of excluding the interests and expectations of these stakeholders is that fundamental choices with respect to evaluation are largely determined by the 'usual suspects', who thus have the opportunity to direct the evaluation before it has actually started.

The results in relation to the influence of the broader context provide a similar picture. Initially it seems that independent evaluations can be implemented without significant problems and that challenges can be solved adequately. However, it is possible that context factors which (may) put pressure on the evaluability, have also influenced this
study (for instance by providing socially desired answers), all the more since this study (as is the case for many evaluations) was performed under considerable time pressure. In other words: it is possible that the absence of big problems can be explained from a focus on the ‘technical’ dimensions of evaluability rather than the political dimensions, which were not explicitly included for several reasons.

The influence of the specific country context on evaluability

Evaluability scores per country do not differ much, which implies that other parameters are probably more important. Only context shows considerable differences, for obvious reasons, but the influence of context on technical evaluability is not of such importance as to be a serious impediment. Certain institutional context can, by the way, have both a positive and a negative effect on various aspects of evaluability. In the case of Rwanda, for instance, the explicit performance culture has a positive effect on the demand for monitoring, the actual development of monitoring systems (mainly focussed on upward accountability), whereas that same focus on performance and ‘quick wins’ may impede the reporting of failures and changes, which at the same time has a negative effect on the learning/feedback function and puts pressure on the evaluability.

The influence of the degree of complexity of the interventions on the evaluability

The evaluability of interventions with a ‘complex’ theory of change (TOC) does not differ substantially from evaluability of interventions with a ‘less complex’ TOC. Interventions with a complex TOC even score a bit higher, possibly because the stakeholders in these interventions invest more in analysis and implementation of M&E systems and practice, and because they – be it justified or not – assume that the interventions involved are harder to finance and that it is more difficult to show their results. Furthermore, the small differences between the two types of intervention can also relate to the fact that certain parameters of complexity (such as the number of stakeholders) in the attribution to the two groups, were not included.

The influence of the funding channel on evaluability

The evaluability scores per funding channel (type Belgian stakeholder) show larger differences than those for the nature of the interventions and the country. Differences are particularly significant between ‘bilateral/NGOs/trade unions’ on the one hand and ‘other stakeholders’ on the other hand, with better scores for the former, even though the latter group also shows examples of good practices. The predominant explanation for this finding could be the less stringent external requirements for the ‘other stakeholders’ as regards intervention plan and implementation practice, which is enhanced by the fact that for part of the stakeholders in this group, ‘development cooperation’ is not a primary task. On the other hand, BTC and NGOs/trade unions as a sector have made a strong effort to improve M&E (particularly monitoring).
5.1.2 Analysis

A comparison between the scores for the five evaluation criteria shows that for all items (including all components and dimensions) the efficiency criterion yields the highest score, followed by effectiveness, sustainability and impact. The score for relevance is situated mainly in the middle. The fact that the same score pattern keeps returning obviously has to do with the degree of difficulty – at all stages of the intervention cycle – of the evaluation of these criteria. Although generalisation is hard in this respect, it can be said that sustainability and more particularly, impact are harder to evaluate than the three other criteria. Impact evaluations require a high level of methodology, while the difficulty of sustainability evaluations often lies in the challenge of making well-founded statements about a situation that will happen in the future.

Other factors may also explain the difference between the scores of the five criteria. First of all, the effects of the control from the donors is tangible mainly at the efficiency level. Donors (and particularly DGD) are first and foremost concerned with correct application of the provided public funds and have imposed important conditions for management of the interventions they provide the funds for. Procedures and formats chiefly follow the donor's priorities, although the stakeholders may add a few specifics via their dialogue with the authorities. Accordingly, the used formats (particularly those for intervention propositions and progress reports) appear to play an important role in determining the evaluability as they may stress, for instance, aspects closely linked to efficiency, but pay little attention to impact. To some extent these formats are also inspired by the desire to create a results-oriented management, for instance through the requirement of linking budgeted expenses to outputs. Furthermore, the high evaluability score for efficiency can be explained from the important efforts made by BTC and NGOs in particular to further develop and implement their M&E systems. These efforts are the result of processes that have been applied within these organisations for some time, but that are also a reaction to the planned screening of NGAs in 2016 which will be examining the quality of the M&E systems.62. The development of M&E systems often involves rather complex initiatives that evolve gradually and in which the changes at intervention level take place rather slowly. These processes – quite logically – follow a bottom-up approach in which the implementation level (inputs – activities – outputs) comes first; incidentally, this is also the level at which the direct usefulness for the organisations involved is most tangible.

However, for several reasons it is not obvious that M&E systems will gradually also 'automatically' integrate the higher levels of the ends-means chain and that the evaluation function in particular will be developed as strongly as the monitoring function. After all: the stimuli and positive preconditions that currently exist for M&E at implementation level (directly demonstrable usefulness; traditional focus on operational aspects; pressure from DGD; relatively minor set-up and implementation requirements) do not or hardly exist for M&E with respect to outcomes and impacts63. An important conclusion that derives from this is that despite the progress of recent years in the development of M&E systems, increasing evaluability – in the first place as regards other criteria than efficiency – is not at all a future certainty if policy and context remain the same.

62 In addition to this screening, a certification of the M&E systems of the stakeholders of the Belgian development cooperation is also foreseen. This certification arises from the DG-D management response to the meta-evaluation of the programmes of NGOs that was also carried out by order of SEO (Meta-evaluatie van de programma’s van niet-gouvernementele actoren, Juli 2013 (Meta evaluation of the programmes of non-governmental stakeholders, July 2013)). The report contains a recommendation with respect to the set-up and implementation of this certification (see chapter 5.2).

63 However, the recently developed ‘Strategy Note on Development Results’ firmly stresses the outcome level and states, for instance (page 3), that outputs are not considered development results. If this Strategy Note is followed up properly and if it is consequently translated into changes in the area of requirements and procedures, it could help to create a turning point.
Another important finding of this study is that irregularities in the intervention design often also have consequences for the implementation and thus affect evaluability directly as well as indirectly. Among other things this applies to:

- The integration of gender and other factors of social differentiation: if no attention is paid to gender at the start (during the analysis of the context and target groups), chances are that none will be paid either during the development of the baseline and, after that, of the M&E system, and that gender will remain underexposed during implementation and hence not included in the monitoring systems and evaluations;
- weak ex ante attention for the impact level (for instance as a result of absence of higher levels or a ‘missing middle’ in the TOC), which extends first into incomplete intervention logic (as specified in the logical framework), then into the proposed M&E system and in the end into the M&E implementation practice, whereby many interventions in fact have no attention for the impact level and do not include impact analysis in their evaluations. As a result interventions or broader programmes may continue for several funding cycles without any reflection taking place about the possible effects of the action, let alone that they be outlined;
- the fact that the characteristics, interests, etc. of stakeholders and other groups involved are insufficiently outlined ex ante implies that they are not or insufficiently involved in the development of M&E systems, in the baseline, and hence in the M&E system in practice, in which they do not play a role either. Further down in the intervention cycle this implies that the chance that interests and expectations of these stakeholders are included in the preparation and implementation of evaluations would be minimal, with negative consequences not only for evaluability, but also for the actual usefulness of the evaluations.
- If the proposition does not describe the way in which the M&E system of the intervention is fine-tuned/related to the national/local M&E system, it comes as no surprise that the M&E system of the intervention in practice is not fine-tuned to that national or local system either.

As illustrated by the above examples, the study has indicated that the quality of the intervention design – or rather: the design phase – is a factor which strongly affects evaluability directly and indirectly. After all: a good design phase is often the first indication of high quality intervention management. Investing in proper intervention preparation appears to pay itself back, whereas initial weaknesses seem less simple to correct at a later stage than one would expect on the outset; in other words: the initial design turns out to be the ultimate reference on which practice is based. How this finding can be harmonised with the current justified aim for administrative simplification will be discussed in the recommendations.

Furthermore, it appears that a weak design phase has different consequences for monitoring and evaluation. Our analysis has shown that, for a number of reasons, from an evaluability perspective the M&E practice scores considerably better than the M&E design. Apparently it is relatively simple to apply corrections at this – mainly operational – level, but shortcomings from the starter phase will continue to have an effect (see above). However, these hiatuses have more severe consequences for the evaluation function because corrections are harder to apply and hiatuses in intervention design and implementation imply that certain important aspects of an intervention, such as effectively reaching initial target groups, the effects of the intervention on different social groups, and the realisation of assumptions and risks, cannot or hardly be evaluated and often are excluded from evaluations altogether. In such cases, the combination of the aforementioned factors can lead to a situation in which ‘independent’ evaluations are in fact being controlled (or at least determined) from within, which ultimately means that gaps and blanks deliberately or accidentally are not acknowledged and controversial subjects are avoided or receive too little attention. In other words, there is a risk that evaluations focus only on reality as it has been defined or interpreted by the intervention involved. The fact that the general usefulness and relevance of evaluations (as an
obligatory part of good development practice) and the desirability of evaluations are seldom questioned could enhance this risk. In such situations even experienced evaluators will have a hard time questioning the focuses of the evaluation where necessary; in many cases they will not even have the chance or be prepared to do so.

Summarising, it appears that the many initiatives aiming at improving the intervention management and thus contributing to more evaluability and (probably) also to better development performance, have accomplished important results. Nevertheless, there are still important shortcomings as well. For instance, the almost total absence of an insight into the impact and sustainability of the effects of interventions that were largely financed with public funds, is problematic from a societal perspective. The same goes for the incompletely developed evaluation function, as a result of which independent evaluations must be carried out in a restrictive framework and for that reason, can but partly achieve their goals in the areas of learning and accountability.

5.2 Recommendations

preliminary remark

In the specifications of this study, point B3 states that the study must be useful for three different groups: the SEO, the DG-D services and the Attachés and partners of the bilateral and non-governmental cooperation. It is the opinion of the research team that each of these groups has an interest and a responsibility in aiming for better evaluability, and it would be best if they addressed the issue from a jointly defined and supported framework and guidelines, each group then working towards the realisation from their own role and specific characteristics. In that way evaluability and evaluability assessment form an additional point of attention in the consultation tradition which exists between the individual stakeholders. As such the strategic recommendations are of importance to everyone. As far as the operational recommendations are concerned, these are important mainly for the partners of the cooperation.

The study team realises that some of the formulated recommendations are quite demanding, at least at first, but maybe not so much in the longer term. As such they can only be applied if the management burden for the stakeholders can be reduced accordingly in a couple of areas. In this framework the efforts to arrive at administrative simplification are a necessary but insufficient condition. It is also important to arrive at a different interpretation with respect to the development of intervention propositions, implementation reports, etc., with a focus on development results (outcomes, impact) instead of on the operational levels (means, activities, outputs). This consideration affects several of the recommendations described below.

5.2.1 Strategic recommendations

1. This study has made an attempt to outline the concept and practice of the assessment of evaluability in the context of the Belgian development cooperation. It showed that the concept and particularly the actual assessment of evaluability are quite unknown in the current cooperation environment. The study further demonstrated that evaluability assessment can be used for several aims which go beyond the ones indicated in the specifications. Assessing evaluability need not necessarily be linked to organising an evaluation. Adequate application of an evaluability test, for instance, could generate important learning effects and contribute to a better practice in that way. More in general, looking at an intervention from an evaluability perspective can contribute in many ways to a better policy and practice in the development cooperation

For these reasons the study recommends that all of the stakeholders at their individual levels, integrate evaluability and evaluability assessment into their operations in a more systematic manner and perceive the two as a means towards better development cooperation. The use of evaluability may not turn into leverage
for bureaucratic control or direction (by the donor; within organisations), but has to form part of change and learning processes which have already been started in many organisations. The aim neither can, nor may be to achieve maximum evaluability; increasing evaluability must be a continuous concern but it has to have an adequate place in a specific context; there will always be a turning point at which the benefits of better evaluability no longer justify the costs.

2. Monitoring and evaluation develop (or have been developed) into good institutionally specified systems with a clear policy and (often) standardised methods. This evolution is positive and an illustration of the belief in the value of evaluation (and monitoring) as a component of good development practice. In this context there is a risk, however, that the value and benefits of evaluation in particular, are no longer questioned ex ante and no longer analysed ex post, which will lead to ritual evaluations or evaluations with limited usefulness or usability, without too many questions asked.

For this reason this study recommends that for each evaluation, explicit action is taken ex ante to analyse and demonstrate the potential benefits of an evaluation rather than to assume implicitly that they are already included in evaluations, so as to arrive at a well-reasoned decision on whether or not an evaluation must be carried out. Introducing a coherent evaluability assessment is an important tool in this process; it can be carried out by the stakeholders with support from the recruited evaluator of the intervention if necessary.

3. The study has revealed important imperfections in the preparatory phase of many interventions: target groups are described and defined superficially (as homogenous groups, without taking into account any elements of social differentiation); important intervention stakeholders are not described or only summarily; many analyses are carried out to account for strategic choices earlier made (whereas those choices should be based on those analyses). These imperfections have important consequences for the quality of the intervention design, for the follow-up of the intervention cycle and also for evaluability.

This study recommends an improvement of the preparatory phase, with a focus on quality rather than on ‘more of the same’. Given that this involves a demanding process, it is important to aim for gradual improvement, which is supported in different ways: via a customised framework (with incentives) from DG-D, via good studies and evaluations that can support the formulation (e.g. at the end of previous phases) and that zoom in on key questions and domains for which insufficient knowledge is available, and via a reduction of administrative requirements and regulations (linked to intervention propositions and reports) that hamper the effectiveness of the development. This administrative simplification preferably coincides with the principles of the recently published ‘Strategy Note on Development Results’, which states that DGD’s results-oriented policy will be aimed at outcomes rather than inputs and outputs.

4. The analysis of the 40 interventions shows that at an operational level (inputs, activities, outputs) much progress has been made in the field of M&E, which contributes to evaluability. However, for several reasons there is little attention for the impact level, whereas the outcome level (coinciding with the specific objective of the interventions) is interpreted in different ways and not always aims at actual changes at target group level. That is problematic: development cooperation essentially is about societal change, but many interventions either cannot demonstrate this type of change or do not specifically aim for it.

For this reason (and in accordance with the Strategy Note on Development Results) the study advocates renewed attention for and upgrading of the outcome and impact levels throughout the entire intervention cycle (intervention design, M&E, etc.). First and foremost this implies a clear definition of these basic concepts and how to apply them. More specifically this study recommends:

- to assure that in intervention propositions, outcomes (the level of the specific objective in the logical framework) are defined as actual changes (in policy,
behaviour, practice) at the level of target groups and local institutions; this involves intermediary results that (via a causal relation) contribute to the desired impact;

• to ensure in intervention propositions that the TOC between the (current) effects and impacts is well developed as it gives attention to, for instance, direct (intermediate) outcomes to which interventions can demonstrably contribute based on a clear theory of change, and to long-term effects at a societal level; such TOC can then become an important instrument for an evaluation which checks to what extent changes that were found (outcomes and impacts) can be assigned to an intervention (or are the result of other influences);

• (in line with international developments) to create a broader definition of efficiency so that outcome and impact are also studied for the relation between costs and benefits.

Obviously this ‘upgrade’ of the outcome and impact level must be extended to the intervention cycle practice and the instruments currently used in the separate phases (intervention propositions, reports, M&E system, etc.).

5. The study showed that, oftentimes as a result of the policy of the implementing organisations, many interventions attach a growing importance to the development of M&E policy and systems and underlying accountability systems for the use of the received funds. Although this is progress in itself, the development and functioning of these systems require a relatively high amount of funds and time from the often already overloaded staff members. For this reason (but also because of the limited attention for the outcome and impact levels - see recommendation 4) it appears that the M&E practice often gets stuck at implementation level: the attention goes to the relatively easily monitored inputs and outputs at the expense of monitoring (more complex) processes of change. In conjunction therewith M&E systems pay too much attention and resources to monitoring at the expense of evaluation. Although this development benefits the evaluability of efficiency, it hampers that of effectiveness and particularly of impact and sustainability.

The study recommends that the implementing partners of the Belgian development work together to create a ‘progressive’ M&E policy, strategy and practice which can/must start with the specification of systems and practices at implementation level (as these already exist in many organisations), but may not be restricted to that. By contrast, it is necessary that these systems and practices gradually develop further and particularly aim for good evaluability of effectiveness, impact and sustainability. Such development (and the corresponding enhancement of the evaluability) is best done gradually, with means, instruments, capacity and experience increasing at simultaneous speed so as to be able to include and integrate more complex functions in due time. Good operational monitoring then provides a solid basis for evaluations which can concentrate on specific questions and aspects which are not included in the monitoring phase.

In the current context it will be important to synchronise this recommendation with the planned screening of NGAs and the harmonisation and certification of the internal monitoring and evaluation systems of the Belgian development stakeholders (see also recommendation 7 below).

6. Taking into account the restrictions with which most Belgian developmental stakeholders are faced, the recommendations suggested above are quite demanding and not easy to realise on several counts. Said restrictions involve both a quantitative lack of funds and a lack of (qualitative) experience and expertise. For this reason it is important to create a framework in which these – ambitious – changes are not only facilitated, but also stimulated and appreciated. DG-D plays a crucial role in this context and in consultation with the other key stakeholders:

• could further revise and simplify the current rules and regulations, instruments and procedures in such a way as to ensure that they focus more on the (intended) development effects and a reduction of the administrative and financial reporting obligations;
• could create incentives to further develop the M&E function (and particularly the evaluation function), which would better allow stakeholders to implement high-quality evaluations which also analyse sustainability and impact; an important initial incentive in this context could be to allow (particularly in new interventions) that baselines are developed at the start of the intervention rather than during the preparatory phase;
• could establish a fund for financing studies and evaluations at effect and impact level, initiated - preferably - by the aggregate Belgian development stakeholders. This fund should finance joint exercises involving several interventions of different stakeholders and implement studies and evaluations which either exceed the means and capacities of the individual stakeholders, and/or are beyond their scope of interest. SEO could take (joint) responsibility for quality assurance, however without being responsible for the management of these evaluations. The function of the fund should also be inspired by the need for more ‘public good’ evaluations of the Belgian cooperation.

7. The planned certification of the M&E systems of the Belgian development stakeholders is an important element in the scope of the findings and recommendations of this study, to which we bring to mind that one of the goals of this study (as stated in the specifications) is to ‘learn lessons that are useful for harmonisation and certification of the evaluation systems of the stakeholders’... It is equally important to point at the planned (2016) screening of the NGAs (with a view to their access to the next funding cycle) in which the M&E system forms one of the screening elements. The legislator has entrusted the implementation of the planned certification to SEO, but SEO is not involved in the screening of the NGAs.

Based on the findings of this study, the following is recommended in this context:
• This study has shown that the quality of M&E systems (in design and implementation) is determined strongly by a combination of other elements of the intervention management and the institutional context and that, conversely, the M&E system affects the quality of those other components of the management. As such the currently foreseen detachment of the certification of these systems from a broader and more integrated approach (as it has been secured in the NGA screening, for instance, at least in first instance) does not seem advisable.
• This study has also shown that DGD has played and could continue to play an important part in the qualitative improvement of the management of the Belgian cooperation (and indirectly in increasing the performance of that cooperation). The findings of this study, which seem to be in keeping with those of a currently performed impact evaluation of NGA actions, would have to ensure that DGD retains the (HR and other) resources to give further meaning to that important role. Whether formal certification is the best approach for implementing that role remains to be seen. The study has established that many stakeholders have already started specifying M&E systems, among other things in view of the planned screening whose methods have been determined via a dialogue between DGD and the stakeholders. It has also established that the progress in specifying M&E systems can differ strongly within organisations (between countries, between different interventions in the same country) and that the local context (e.g. in fragile states) can be a substantial determinant factor in this respect. This essentially involves qualitative processes in which context is an important parameter and the gradualness and set-up of ownership (have to) take a central position. Chances are that a certification approach, which – unavoidably – must be implemented quite formally and in a standardised way and which first and foremost is going to provide a ‘snapshot’ of the situation at a certain moment (without any connection to the process underway), will interfere with the processes running at the moment and in the end may generate counterproductive effects from an evaluability point of view. That does not mean that DGD should leave further specification of M&E systems totally to the stakeholders. If it so desires, DGD must have access to the results of the...
running M&E processes. For this purpose it could, for instance, provide the means for specification and application (by the stakeholders or with external guidance) of a diagnostics instrument which would allow the stakeholders (and DGD) to get an insight into the strengths and weaknesses of their M&E system, which allows them to develop a tailor-made results-oriented action plan for the purpose of improving their M&E system and practice, in which the method and time span for adjustment of specific components of the M&E system and the practice will be indicated and in which stakeholders are asked to report their progress to DGD.

- In connection with the previous point and in case one opts for implementing the certification in accordance with the current plan, what has to be avoided is that the certification grows into a pure meta-evaluation approach of M&E in the sector. Furthermore it is important to realise that even if one goes ahead with the certification of M&E systems, it will remain important to perform sufficient and sizeable external evaluations of the quality of the work of the development stakeholders in the future (in addition to the evaluations initiated by the stakeholders themselves – see also the operational recommendations below). However, other instruments are available for this (or must be made available – see recommendations 6 and 11).

### 5.2.2 Operational recommendations

A few good practices can be derived from the analysis of the factors that affect the evaluability (see chapters 3 and 4) which not only benefit the evaluability, but the development performance in general. We will limit ourselves to a few important operational recommendations which often have implications in the different phases of the intervention cycle and/or are of considerable importance in our opinion. Some of these recommendations partly overlap strategic recommendations described above.

#### 8. Improvements of the intervention design

As indicated earlier (see recommendation 3) these improvements should preferably be interpreted as a gradual process. Important initial steps (insofar as necessary) include:

- a better description (differentiation) and definition of direct target groups;
- the development (or improvement) of an adequately specified baseline in which (1) the results of the previous point are included and (2) a definition and gradual completion (‘upward’) of the theory of change takes place and, accordingly, (3) the formulation of the intended effects (specific objective) and impact (including corresponding indicators) and crucial hypotheses is refined and/or completed, and (4) baseline data are collected with respect to the selected indicators, etc.

#### 9. Completion of the M&E policy and its translation into practice

An increasing number of organisations is working towards developing an M&E policy. However, in many cases this policy and particularly its application shows a number of imperfections which can differ greatly between organisations. The recommendations below relate to shortcomings that occur quite often but not necessarily for all stakeholders.

In developing and implementing an M&E policy it is important to:

- stress the development of the evaluation function (in a strict sense) – including the relation to monitoring;
- pay more attention to the implementation of the policy at a country and intervention level, while including the local elements and situations (involving the local partners in the formation of the local policy and the local practice; harmonisation of own M&E policy and that of local stakeholders);
- develop a clear plan of action in which M&E efforts and results are used for actual analysis, learning and change.
10. **Development of the M&E function at intervention level**

In addition to the necessity of increased attention for the ‘higher levels’ in the end-means chain and for evaluation (see recommendations 4 and 5) it is important that:

- the set-up and functioning of M&E systems are planned well; this does not necessarily imply that these systems have to be developed during the formulation of the intervention, but it does mean that their set-up, development and functioning must be properly defined and planned with ample attention for the required personnel and financial resources, capacity increase, etc.;
- the different M&E components are actually combined into a coherent system which comprises the entire M&E field, avoids overlap between the individual fundamental components, and implies a good definition of the required resources (quantitative and qualitative) and the distribution of tasks (the role of the different stakeholders);
- the role of the different stakeholders in M&E systems is optimised, while avoiding that M&E becomes the monopoly of the partners with responsibility for the implementation and instead checking, based on the principle of subsidiarity, who can contribute what, where and how to M&E;
- sufficient attention is given to providing feedback of the monitoring results to all of the stakeholders, in order to keep feeding awareness for the purpose of (good) data collection and contributing to the sustainability of the system;
- a check is performed on how the M&E system of the intervention relates to local and national M&E systems and, where necessary and relevant, which role the intervention can play to set up, improve and secure those systems.

11. **Preparation of evaluations and implementation based on constant focus on their final use**

This recommendation builds on (among other things) recommendation 2 and starts from the premise that first and foremost the quality of an evaluation depends on its effective use, which is largely determined by its usefulness and usability. A thorough reflection on the future benefits of the evaluation (recommendation 2) is a necessary yet insufficient condition in this context.

In order to optimise the final use of evaluations it is important to:

- in accordance with recommendation 2, provide for an evaluability assessment in each evaluation, which, among other things, will analyse and demonstrate the potential benefits of a planned evaluation;
- place the planning and implementation of an evaluation in a ‘portfolio approach’ in which not all of the interventions need to be evaluated in all of their aspects, but instead strategic choices are made in keeping with the available resources, requirements of accountability, environment factors, etc.

It could also be interesting to distribute the interventions over several groups in line with the current insights into the functioning and impact of the intervention:

- interventions for which previous evaluations and studies have shown at length that they have an impact (often interventions with a less complex TOC), do not have to be evaluated for impact time and again. Therefore, a counterfactual need not be specified for this group of interventions; it would suffice to evaluate effectiveness and efficiency;
- interventions that are known to be well-implemented and are candidates for repetition or substantial expansion, but whose potential for generating impact is insufficiently known, should be subjected to a thorough impact evaluation. For these interventions it is necessary that they can be evaluated effectively for their impact and that this is taken into account as from the design of the intervention. This implies that the different aspects of evaluability of impact which were examined throughout the study, indeed get attention (among other things TOC up to the level of impact; identification of crucial elements up to the level of impact; translating TOC in M&E system up to the...
level of impact; specification of a counterfactual; good baseline; etc.). Particularly the presence of data for a counterfactual\textsuperscript{64} and data on external assumptions is important before proceeding towards impact evaluation\textsuperscript{65};

- Interventions of which it is insufficiently known whether the chosen implementation method is the best, can best experiment with different methods during the implementation and go for evaluation which compares the functioning and effectiveness of these different methods (in this context, see the idea of \textit{structured experiential learning}, footnote 28);

- involve the end users of the evaluation (stakeholders and others involved in the intervention) in the evaluation as early as its preparation (e.g. by considering their interests and expectations);

- define terms of reference with well-defined and well-thought out main questions, which also allow (or even stimulate) evaluators to perform research outside the strict framework of the evaluation objectives and main questions.

\textsuperscript{64} This does not necessarily imply the performance of an RCT. Chapter 2 described several options for identifying a counterfactual.

\textsuperscript{65} While the absence of an explicit TOC up to the level of impact can be remedied in the course of an impact evaluation (because people responsible at interventions often have an implicit TOC without putting it to paper), the absence of an M&E system in practice which collects valid data for the intervention-group/situation and a counterfactual is harder to remedy during an impact evaluation.
Annexes

(in separate documents)

1. Special specifications
2. Description of the research methodology and approach
3. Study framework
4. List of 40 analysed interventions
5. List of the most important consulted documents
6. List of the most important persons contacted
7. Explanation to the statistical analyses and SWOT analysis
8. Detailed score overview
To evaluate in a credible and meaningful way: between dream and reality

A study of the evaluability of (co)-financed interventions of the Belgian Cooperation