This item is the archived peer-reviewed author-version of:

To evaluate or not : evaluability study of 40 interventions of Belgian development cooperation

Reference:

Holvoet Nathalie, Van Esbroeck Dirk, Inberg Liesbeth, Popelier Lisa, Peeters Bob, Verhofstadt Ellen.- To evaluate or not : evaluability study of 40 interventions of Belgian development cooperation
Evaluation and program planning - ISSN 0149-7189 - 67(2018), p. 189-199
Full text (Publisher's DOI): <https://doi.org/10.1016/J.EVALPROGPLAN.2017.12.005>
To cite this reference: <https://hdl.handle.net/10067/1487660151162165141>

This item is the archived peer-reviewed author-version of

To evaluate or not: Evaluability study of 40 interventions of Belgian development cooperation

Reference :

Holvoet, N., Van Esbroeck, D., Inberg, L., Popelier, L., Peeters, B. and E. Verhofstadt (2018), "To evaluate or not: Evaluability study of 40 interventions of Belgian development cooperation", Evaluation and Program Planning 67: 189-199, <https://doi.org/10.1016/j.evalprogplan.2017.12.005>

To evaluate or not: evaluability study of 40 interventions of Belgian Development Cooperation

1. Introduction

Within development cooperation, evaluations have become increasingly important to account for the use of resources, learn from experiences of the past and support future policy-making and practice. At the same time, there is a growing concern with interventions' evaluability which most development agencies define in line with the Organisation for Economic Cooperation and Development/Development Assistance Committee (OECD/DAC) as *"the extent to which an activity or project can be evaluated in a reliable and credible fashion"* (OECD/DAC, 2002:21).

Poor evaluability is generally related to a number of challenges in the area of the intervention and evaluation design, data availability, as well as the stakeholders' and the broader context readiness for evaluation and the use of evaluative findings (Davies, 2013; Wholey et al., 2010). While these challenges generally affect an intervention's evaluability, they are particularly pressing in the context of development cooperation where interventions tend to be more complex, with a multitude of stakeholders holding competing interests. Interventions are also often implemented in settings characterised by a scarcity of financial and human (M&E) resources and environments that are not always conducive to independent evaluations and evidence-based policy making and practice.

Against this background, evaluability assessments (EA) which analyse whether a number of basic conditions are fulfilled for sound, relevant and useful evaluation have been on the rise within development agencies over the past decade (Davies and Payne, 2015). Besides a cost-effective pre-evaluation mechanism that provides an important input into an evaluation's desirability, objectives, timing and methods, the application of EA has also gradually broadened over time. Somehow fuelled by the growing acknowledgement of the usefulness of evaluative tools and skills (see e.g. widespread use of 'theory of change') prior to and during the intervention cycle, EA is currently also being used

to improve the intervention design, feed into the design and the development of an M&E system and contribute to the usability and effective use of the evaluative outputs (Peersman et al., 2015).

Also within Belgian development cooperation, and particularly in the Special Evaluation Office (SEO), the concept of evaluability has recently gained more interest, especially as experiences with external evaluations suggest that the initial conditions to evaluate are often not or only partly satisfied.

Against this background, SEO commissioned ‘an evaluability review’ in 2015 to diagnose and analyse the current evaluability of different types of interventions of Belgian development cooperation (SEO, 2015). This article presents and discusses the main findings of the EA¹. In the next sections we first describe the study framework, sample and the methods of data collection and analysis.

2. Methodology

2.1 Evaluability Assessment Framework

While development agencies have started to apply evaluability assessment since the turn of the century, so far, there does not exist a standardised framework (Davies and Payne, 2015). Our study framework draws upon earlier contributions to the EA field while taking into account a number of considerations specific to the study’s context.

‘Evaluability assessment’ was first introduced by Joseph Wholey (1979) in the early seventies and was initially conceptualised as a mechanism to counterbalance poor quality impact evaluations which were not delivering the expected outcomes. In many cases, the (often substantial) cost of such impact evaluations could have been prevented as some basic conditions for sound evaluations were often not fulfilled (Leviton et al., 2010). In Wholey’ view, these conditions include amongst others realistic and agreed upon program goals, well defined information needs, data availability and the

¹ For a full overview of findings, see Authors (2016).

willingness and ability to use the evaluation outputs (Wholey et al., 2010). Evaluation assessment is then used to diagnose whether these initial conditions are fulfilled, whether it is desirable to proceed with impact evaluation and/or which adjustments can be made in order to make the evaluation credible and useful.

As discussed in Smith (1990) and Trevisan (2007), over time the scope of EA has broadened from summative impact evaluations to formative and process evaluations. Reviews of EA practice (see e.g. Davies, 2013; Trevisan, 2007) highlight that EA is currently also used to assist management in establishing the pre-conditions for a comprehensive evaluation (see Schmidt et al., 1982), to agree upon realistic program goals, evaluation criteria and expected use of evaluation information (Wholey et al., 2010; McKinney, 2010). In addition, evaluability assessments could contribute in strengthening and monitoring of M&E systems (Dahler-Larsen, 2013) and in directing financial resources to those evaluations that could reveal important and useful information (Leviton et al., 2010; McKinney, 2010).

Given the broad spectrum of EA goals, and diverging opinions on the necessary conditions of a reliable, credible and useful evaluation, the lack of a standardised EA framework should not come as a surprise. In the development of our own framework and the EA process we have been inspired by EA reviews (see e.g. Davies, 2013; Trevisan, 2007) while taking into account a number of factors specifically related to the study's context.

In line with procedures for evaluations commissioned by SEO, an advisory board was installed consisting of the evaluation experts from different types of implementing agencies including Belgian Development Agency (BTC) (bilateral development cooperation), actors of indirect development cooperation (NGOs, trade unions, universities, cities, etc.), Flemish and French speaking federations of NGOs, etc. At different stages of the EA, the advisory board was informed and provided useful feedback and input into the study's design and implementation which also stimulated the

acceptability and use of its findings (see also Thurston and Potvin, 2003). Importantly, at the outset of the EA there was a considerable amount of resistance against the study. This was partly related to unfamiliarity with the notion of 'evaluability', the initial aim of the study which was 'to make all interventions evaluable within the near future' (SEO, 2015) as well as the implicit and underlying drive towards more 'golden standard' impact evaluation designs. In order to counter the resistance, the study team strongly invested in unpacking the notion and outlining a number principles which were accepted by all main stakeholders, including the SEO. More specifically, we have approached 'evaluability' as a continuum assuming that no single intervention is evaluable in all its aspects while we neither believe that interventions exist that cannot be evaluated at all. Second, we have clearly distinguished among 'evaluability' and 'inherent (development) quality' as many interventions which are highly relevant from a development perspective might face particular challenges in terms of evaluability. This may be due to more complex Theory of Change (ToC), less consensus among stakeholders about intervention objectives and design, lack of Specific- Measurable- Attainable – Relevant – Time Bound (SMART) indicators, lower context readiness, etc. Third, we have differentiated among the classic DAC criteria of relevance, effectiveness, efficiency, impact and sustainability as not all criteria imply the same evaluability conditions. While these principles were key to the elaboration of the EA framework and process, they also segued into an adjustment of the initial EA's aim, broadening it towards EA's use as a formative evaluation tool and instrument of M&E capacity building.

In line with most EA discussed in Davies' review (2013), our study framework distinguishes among three dimensions, i.e. 1) analysis of the intervention design, including the underlying Theory of Change (theoretical evaluability); 2) analysis of the evaluability in practice, including availability of information regarding the implementation and results of the intervention as well as the M&E system in practice (practical evaluability) and 3) analysis of the evaluation context, focusing on its conduciveness to evaluation. This broadly relates to 'politics of M&E', referring to the influence of the organisational and broader institutional context with its differently positioned stakeholders on

different dimensions of evaluation. Compared to Davies’ study, especially the second and most important dimension of our EA is more detailed, while also the third part is slightly more elaborate. This results from previous research on politics of M&E (Author et al., 2012), and from discussions and reflections within the team and with the advisory board.

In a next stage, each of these three key dimensions was further disaggregated into a number of components and further made operational into specific items, differentiating as well among the different OECD/DAC criteria². Such operationalisation into a kind of checklist was considered particularly useful for a number of reasons. First, making explicit what each of the evaluators in the EA team considered important conditions under each of the dimensions and components was necessary to structure the EA and to find a consensus within the team. Second, the use of a checklist allowed to be transparent towards the broad range of stakeholders involved about the items that would be assessed and it was also a basis for discussion during the advisory board meetings. Third, while it was not the initial aim of the checklist, at the end of the study various stakeholders highlighted that the checklist was particularly useful in their evaluation capacity building efforts.

Table 1 gives an overview of the three key dimensions, the eight main components and the maximum number of items scored under each component. The number of items scored for the different OECD/DAC criteria are displayed in each of the columns. A full overview of the framework with the different components (2-digit level) and the items (3-digit level) is added in Appendix A.

Table 1. Overview of the review framework

DAC Evaluation criteria	Relevance	Effectiveness	Efficiency	Impact	Sustainability
Components (number of items)	(number of items scored)				
1. Analysis of the intervention design (theoretical evaluability)					
1.1 Underlying analysis (7)	(7)	(4)	(0)	(0)	(0)
1.2 The intervention logic and the theory of change (ToC) (8)	(1)	(6)	(5)	(5)	(3)
1.3 The proposed M&E system (9)	(5)	(7)	(9)	(7)	(7)

² We use the OECD/DAC definitions of the different OECD/DAC criteria. See OECD/ DAC (2002).

1.4. Consistency and adaptation of the intervention logic and ToC (3)	(3)	(3)	(3)	(3)	(3)
2. Practice regarding intervention implementation, intervention management and context (practical evaluability)					
2.1 Basic information regarding intervention implementation (11)	(5)	(10)	(11)	(10)	(8)
2.2 The M&E system in practice (12)	(11)	(12)	(12)	(12)	(12)
3. The evaluation context					
3.1 Attitude of key players (9)	(9)	(9)	(9)	(9)	(9)
3.2 The broader context (3)	(3)	(3)	(3)	(3)	(3)
Aggregate score (number of items scored)	(44)	(54)	(52)	(49)	(45)

2.2 Sample

Our evaluability review draws upon data collection for a sample of 40 project interventions in four countries: Benin, Democratic Republic of Congo, Rwanda and Belgium. The sample was purposively selected and is composed of interventions with considerable expenditures in 2014. In order to check for systematic differences among interventions funded through different channels of development cooperation, our sample includes interventions financed through the Belgian Development Agency (BTC), non-governmental organisations (NGO), and other actors such as universities, cities, etc. The underlying rationale is that these different groups of actors have different relationships with the Directorate General of Development Cooperation (DGD) with different requirements in terms of M&E and repercussions on ‘evaluability’ (see also Lloyd. et al., 2013). Drawing upon the assumption that also the complexity of the ToC influences ‘evaluability’, we included interventions with either complex or less complex ToC³. As an initial proxy for ToC complexity, we used ‘sectors’ which later on had to be adjusted as it turned out to be a useless proxy⁴. Finally we arrived at a sample composed of

³ Differences between countries, complex/less complex theories of change, and intervention channels are the subject of a separate paper. See also section 4.3.

⁴ We initially considered ‘education’, ‘health’, ‘water and sanitation’, ‘social economy’ as sectors with less complex ToC and ‘climate and environment’, ‘conflict and peace’, ‘governance’, ‘agriculture and rural development’ as sectors with more complex ToC. After the initial selection and review of the project documents we had to refine this classification as different projects (eg. institutional strengthening projects) in ‘less complex’ sectors turned out to have more complex ToC (and reversely).

10 interventions in each of the four countries, 60% complex versus 40% less complex ToC, 25% bilateral interventions, 50% interventions designed and implemented by NGOs and trade unions and 25% interventions carried out via other stakeholders (universities, cities, Flemish Association for Development Cooperation and Technical Assistance/Association for the Promotion of Education and Training Abroad-VVOB/APEFE, Belgian Investment Company for Developing Countries-BIO).

2.3 Data collection

The evaluability review draws upon a combination of secondary and primary data collection. We set out with a round of secondary data collection of documents provided by the implementing agencies which included (if relevant and available) the intervention proposal, baseline report, progress reports (e.g. midterm, yearly reports and end term reports), studies and evaluations and M&E policy documents of the organisations involved. The collection and analysis of documents has been restricted to this same set of basic documents and took approximately 2 days per intervention.

Primary data was collected from April to June 2015 through focus group discussions at the organisations' headquarters (Brussels) and through interviews with important key stakeholders on the ground. Focus group discussions took a total of 3 days and were organised in four different sub-groups: NGO and other actors (Dutch speaking) (8 participants), NGO and other actors (French speaking) (5 participants), BTC (7 participants), DGD (4 participants). Subsequently, each of the countries (including Belgium) were visited for three weeks⁵, using approximately 2.5 days per intervention during which 15 to 25 persons were interviewed (individually and/or in small groups). These site visits and interviews were particularly useful to get a better grasp of local intervention and evaluative realities, the degree to which an M&E system was actually in place and functioning as well as the contextual reality (see also Wholey et al., 2010). Getting insight into the latter dimension also

⁵ Rwanda was visited from April 22nd to May 13th; Benin from the 1st to the 19th of June; DRC from the 8th to the 26th of June; interventions in Belgium were visited between the 1st and 26th of June.

benefitted from our previous research in the specific country settings as well as the involvement of local evaluators with locally grounded knowledge in each of our country teams. Some of the key stakeholders have been interviewed together, either because they were actors with comparable M&E responsibilities or positions within the intervention, or because of time constraints. After having interviewed the most important stakeholders, the members of the review team organised a confidential feedback session in which the first preliminary findings were discussed and verified with the key stakeholders. This allowed to validate and fine-tune findings and to add additional data where needed.

2.4. Scoring and data analysis

The eight components and the 62 items of the review framework have been scored with regard to (some of) the OECD/DAC criteria using scores that vary between 1 (weak performance) and 5 (strong performance). In order to reduce subjectivity, scoring has been done by (at least) two persons on the basis of a grid which explains for each component, item and specific OECD/DAC criteria the different scores (see also IADB, 2002, 2007; ILO, 2012; Poate et al., 2000). In case of differences between the scores assigned by the two evaluators, the scores were calibrated in order to arrive at a final score. Scoring for the theoretical evaluability items mainly (but not exclusively) draws upon secondary data while primary data was particularly useful for scoring of the other dimensions.

Scoring of the eight components (2-digit level) can be done in two ways: firstly a score can be assigned on the basis of the evaluation grid, disregarding the specific scores on the different items that are related to the component; secondly an average can be calculated based on the item scores related to a specific component (e.g. scoring of 2.1. is based upon scores for 2.1.1. to 2.1.11). In calculating the average, different weighting procedures may be used which are associated with specific working assumptions: an un-weighted aggregation of different items for instance implies that all items related to a component are equally important for evaluability, which is however not

necessarily the case (see Davies, 2013; Peersman et al., 2015). In the end, we decided to use the direct scoring at component level. Some caution is warranted in interpreting scores: these cannot be interpreted as absolute numbers and are only used for comparative purposes. They particularly allow to get a quick overview of those components, items, and OECD/DAC criteria that score relatively good or poor in terms of evaluability.

Based on the scoring for the 40 different interventions, indexes have been calculated for each of the items and components, differentiated alongside the different OECD/DAC criteria (where applicable). For example, item 1.2.7 ('The external assumptions are clearly identified and explored/estimated') was considered important for the evaluability of all DAC criteria and thus interventions were five times scored on this item. Subsequently, indexes were calculated for all different items, each time differentiated alongside the relevant OECD/DAC criteria⁶. The 'relevance evaluability' index for item 1.2.7 was e.g. calculated as follows: [(number of interventions with score 1 for relevance x score 1) + (number of interventions with score 2 for relevance x 2) + (number of interventions with score 3 for relevance x 3) + (number of interventions with score 4 for relevance x 4) + (number of interventions with score 5 for relevance x 5)]/maximum score (=number of interventions scored x 5). In addition, an overall evaluability index, based on all criteria, has been calculated for each component and item⁷ (see Appendix A).

While the 40 interventions could often be scored on all items, some 'non-applicable' scores were assigned. This was amongst others the case for component 1.4 related to 'consistency and adaptation of the intervention logic and ToC' which was only scored in case changes effectively took place. Based on the scores, we have also ranked for each of the different OECD/DAC criteria the different items. More specifically, the 44 items that have been scored on 'relevance' have been ranked from 1 (highest score) to 44 (lowest score). A similar procedure has been followed for

⁶ In calculating these indexes the working assumption was that all interventions have an equal weighting.

⁷ In calculating these indexes the working assumptions was that all criteria and all interventions have an equal weighting.

effectiveness (ranking from 1 to 53), efficiency (ranking from 1 to 51), impact (ranking from 1 to 49) and sustainability (ranking from 1 to 44). We have as well looked at item ranking within each of the three dimensions (theoretical evaluability, practical evaluability and context) while we also ranked the 8 components, differentiated for each of the different OECD/DAC criteria in order to distil those components on which the 40 interventions generally score better (or poor) in terms of evaluability.

3. Findings and analysis

In what follows we provide a general overview of the findings at the level of the eight components of the review framework, differentiated alongside the five different OECD/DAC criteria. Table 2 presents an overview of the average scores per OECD/DAC criterion and the evaluability index for the different components (the item scores are provided in Appendix A). To ease interpretation, we have also converted raw scores to simple percentages of the maximum possible score (see Table 2 and Appendix A). We visualised our findings by regrouping index-scores into four different categories and assigned to each category a different shade of grey. This also allows comparability throughout the analysis.

Table 2 : Average scores^(*) per DAC criterion and evaluability index for the 40 interventions

Components (°°)	Relevance	Effectiveness	Efficiency	Impact	Sustainability	Evaluability index
Dimension 1: THEORETICAL EVALUABILITY						
1. 1 Underlying analysis (7)	3.65 (66.25%)	3.60 (65%)				3.63 (65.75%)
1.2 The intervention logic and the theory of change (ToC) (8)		3.00 (50%)	3.50 (62.5%)	2.10 (27.5%)	3.25 (56.25%)	2.96 (49%)
1.3 The proposed M&E system (9)	2.45 (36.25%)	2.83 (45.75%)	2.85 (46.25%)	1.95 (23.75%)	2.46 (36.5%)	2.51 (37.75%)
1.4 Consistency and adaptation of the intervention logic and ToC (3)	3.55 (63.75%)	3.44 (61%)	3.52 (63%)	3.50 (62.5%)	3.75 (68.75%)	3.54 (63.5%)
Dimension 2: PRACTICAL EVALUABILITY						
2.1. Basic information regarding intervention implementation (11)	2.95 (48.75%)	3.20 (55%)	3.53 (63.25%)	2.13 (28.25%)	2.62 (40.5%)	2.88 (47%)

2.2. The M&E system in practice (12)	3.03 (50.75%)	3.3 (57.5%)	3.50 (62.5%)	2.53 (38.25%)	2.85 (46.25%)	3.04 (51%)
Dimension 3: EVALUATION CONTEXT						
3.1. Attitude of key players (9)	3.68 (67%)	3.70 (67.5%)	3.70 (67.5%)	3.58 (64.5%)	3.63 (65.75%)	3.66 (66.5%)
3.2. The broader context (3)	4.10 (77.5%)	4.25 (81.25%)	4.25 (81.25%)	4.10 (77.5%)	4.15 (78.75%)	4.17 (79.25%)

(°) raw scores between 1 and 5 and converted into simple percentages of the maximum possible score; (°°) number of items between brackets

Index 4.01 – 5.00 (75.1%-100%)
Index 3.01 – 4.00 (50.1%-75%)
Index 2.01 – 3.00 (25.1%-50%)
Index 1.01 – 2.00 (0-25%)

Table 2 shows that the context dimension scores remarkably better than the theoretical and practical evaluability in particular. Comparing component level scores indicates that the ‘broader context’ is outperforming the others, while the ‘proposed M&E system’ is overall the weakest scoring component. Except for components and items under dimension 3, the maximum scores are around 3.5 with 60% of the 244 items scoring below the average level (Appendix A). Focusing on different OECD/DAC criteria shows that ‘impact’ has for almost all components the weakest scores, whereas effectiveness and efficiency show relatively high scores. The overall weakest score is the ‘proposed M&E system’ at impact level.

In the remaining sections we subsequently zoom into the main findings at the level of the three dimensions of the assessment framework.

3.1 Theoretical evaluability

Underlying analysis. Scores for this first (out of four) component are relatively high. The quality of analysis strongly influences the evaluability of interventions, as without a clear identification of the development problem and beneficiaries and a clear link between the analysis and the expected

results, the relevance and effectiveness of the intervention can hardly be evaluated (see also Inter-American Development Bank, 2007; International Labour Organization, 2012).

Among the seven items included under this component, the integration of a gender analysis is the weakest. This is largely in line with the findings of a recent evaluation on the gender sensitivity of Belgian development cooperation (see Caubergs et al., 2014). The absence of a gender analysis at the outset of an intervention may substantially lower evaluative quality: gender relations often have an effect on the implementation and results of intervention and failing to integrate a gender dimension in the analysis risks to leave unveiled possible reasons for low performance at the moment of the evaluation. This holds more generally for all dimensions of social differentiation in the target group which tend to be underexplored in the initial analysis at the moment of the intervention design in more than half of the 40 interventions under study.

Intervention logic and Theory of Change (ToC). This second component influences the evaluability of all five OECD/DAC criteria. If, for example, the intervention logic and ToC is well elaborated until the output level (implementation theory), but not until the impact level, the evaluability of effectiveness will be stimulated, while the evaluability of impact will be a challenge. In fact, the elaboration of the intervention's impact theory is one of the weakest elements for many interventions. Many interventions focus especially on the more operational implementation level and the direct effects of the interventions as formulated in the specific objectives. However, even if effects at the impact level are not immediately visible, the elaboration of the theory of change until the impact level also provides insights in the implementation modalities that are most effective and efficient, also increasing the possibilities to realise impact.

'The proposed M&E system' focuses on the M&E system that is described in the relevant intervention documents. Overall and for all five OECD/DAC criteria, 'the proposed M&E system' has the weakest score with particularly low scores for those items (e.g. gender, higher levels in the ToC and internal

risks and external assumptions) that showed already deficiencies in the analysis and ToC. While this holds true for all countries and all implementation channels, we also came across some good practices (see Box 1).

Box 1. Description of risk monitoring in a food security project implemented by the NGO Caritas in Rwanda

The project documents of PASAB II (*Projet d'Appui à la Sécurité Alimentaire au Bugesera*), an intervention of the NGO Caritas in Rwanda, highlight risks for each of the intended outcomes which could affect their implementation and realisation. A risk analysis framework is included 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 materialize. Besides the follow-up of the different intended outcomes in a format which provides room for adding qualitative comments, the risks are also monitored through a specific format which compares the expected risks with the actual situation on the ground. These documents contain useful information that can also be used in evaluations, because they provide an insight into possible underlying reasons for the success or failure of interventions.

As there is also a strong correlation between the proposed and the actual M&E system, the scores for the elaboration of the M&E system in the initial phase are a good indicator for the practice in a later phase, although weaknesses in the design can sometimes be corrected for later (see below). The weak scores on this component can to a large extent be explained by the fact that real attention for M&E often only increases after take-off of the intervention which is somehow stimulated by the fact that the requirements of the Belgian donor with respect to the description of the M&E system in the intervention proposal are rather vague.

The last component within this dimension is '*consistency and adaptation of the intervention logic and theory of change*'. This component is important for the evaluability as a lack of reporting of adaptations in the intervention logic and theory of change and lack of adjustments in the M&E system lead to missing information concerning the real situation, which is likely to result in erroneous conclusions as regards the intervention's quality and the factors that influence the implementation and impact of interventions. While the total score on this component is relatively high, the scores on the three items that fall within this component vary considerably, with the highest scores for reporting and justifying changes in the intervention logic and ToC and the lowest scores for available

information on the vision and opinions regarding these changes. The fact that about half of the eleven interventions which experienced substantial changes during the intervention did not adequately update their M&E systems is somehow alarming. Yet, our findings are not unique in this regard as Thurston and Potvin (2003) reported earlier that it is not uncommon for programmes to evolve and make substantial changes without translating this into program documents and systems.

3.2 Practical Evaluability

Dimension two consists of two components, i.e. ‘availability of basic information’ and ‘the M&E system in practice’. *Data availability* is an important problem for many of our interventions and this is in line with findings from earlier review studies (see e.g. Davies and Payne, 2015; Poate et al., 2000). Our EA shows that data availability is especially a problem for those issues neglected in the initial analysis (e.g. gender and social differentiation), for higher levels of the ToC and for items outside the intervention (e.g. counterfactual, external assumptions). Especially baseline information on a counterfactual is mostly lacking. Weak scores on this item could be explained by the fact that the identification of a counterfactual is not easy⁸ and requires some methodological knowledge. Moreover, the use of a counterfactual is not yet common within Belgian Development Cooperation and is often conceived in an overly narrow way. While it is obvious that the first best option of ‘a randomised controlled trial-type of control group’ is often not feasible, or even not desirable, in the interventions under study, many second-best options that proved to be feasible without too much additional effort were disregarded⁹.

⁸ For instance because an intervention is often not implemented in isolation and may be intertwined with other interventions; because some interventions are targeted at the entire population which leaves little room for the selection of a similar non-intervention group, etc. However, in all these cases and with the necessary methodological knowledge (and creativity) it is possible to identify valuable counterfactuals. See also footnote 9.

⁹ Depending on the availability of resources, other options include e.g. selection of non-intervention group through matching, selection of a similar group that receives another intervention (this allows to evaluate *differential* impact), use of statistical controls, generic controls or some more complex types of before-after such as panel or time series (see Rossi et al, 2004; Bamberger et al, 2004).

Scores for 'the M&E system in practice' are relatively high and generally higher than for 'the proposed M&E system'. This particularly holds for the 'monitoring' component of the system and to a much lesser extent for the more 'evaluative' activities. In some of the cases where 'monitoring' systems were not clearly described on paper, they tend to function better on the ground than expected. On the other hand, the 'evaluative' component of M&E systems is in all interventions much less developed in terms of e.g. vision, clear allocation of responsibilities, time, expertise and institutional arrangements. It then does not come as a surprise that M&E systems are mainly geared towards accountability rather than learning as it is particularly the learning objective that depends to a much higher extent on the analytical (evaluative) component of the system. The existing M&E systems are overall less well apt to provide inputs for sound impact evaluation and evaluation of 'relevance'. Box 2 provides information on a good practice case which, exceptionally, also scores high on efforts of alignment and institutional strengthening of national/local M&E systems.

Box 2. The development of a cross-intervention localized M&E approach with special emphasis on the learning dimension (BTC Katanga)

The EDUKAT project (carried out by BTC in the DRC) contains clear indications with respect to the set-up and organisation of both angels (monitoring and evaluation) of an M&E system. It is a localized version of the general policy and instructions developed by BTC. 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 into a coherent whole. The work plan defines the steps for developing a baseline, including a follow-up matrix, 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 based on the commitment to guarantee learning and continuity of the M&E system within a long-term (12 years) engagement. In this way also 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 is taking shape.

Notwithstanding some exceptions (see box 2), one of weaker issues of many intervention's M&E system is 'the alignment to the national M&E system' (see also Lloyd et al., 2013). Especially interventions which are financed by non-governmental organisations often do not even consider aligning to or strengthening these national systems. While this might be unrealistic for many of these interventions and even lower EA in the short run if such national systems have serious deficiencies in

terms of capacity and independency, it might be crucial, from a sustainability perspective, to explore if an intervention's M&E system, or at least some of the key indicators, could gradually be aligned to the national system. Using parallel M&E systems might seem optimal in the short run, yet, it blocks national M&E systems from maturing and undermines local embeddedness and use/influence of evaluative findings which in the long run also jeopardizes the quality of an intervention's M&E system and evaluability (source deleted for blind review).

3.3 The evaluation context

The last dimension, the evaluation context, consists of two components, the 'attitude of key players' and 'the broader context'. Our study findings point at a rather positive *attitude of key players*. This is essential for an intervention's evaluability as a negative attitude makes the work of evaluators in all circumstances very difficult, even with high scores on the theoretical and practical evaluability.

Whereas impact and sustainability score somewhat lower than the other three criteria differences are rather minimal. The identification of the principal users of evaluations and their expectations is the item with the lowest scores. As there tends to be some trade-off among different objectives of evaluations, with different types of evaluations emphasizing different principles and methodologies (e.g. accountability-focused evaluations need external validity and representativeness whereas learning-focused evaluations rather focus on some specific dimensions and benefit from a higher degree of internal validity), it may jeopardize an evaluation's quality when the principal users' needs are not clearly identified or when too many objectives are to be reached and too many OECD/DAC criteria are to be covered by one single evaluation.

The last component, '*the broader context*', focuses on the context beyond the intervention boundaries. While there is some variation among the different countries in our sample, 'the broader context' is overall the strongest component of the review framework, with also the three items within this component scoring in the highest index category. These high scores illustrate the fact that independent evaluations could take place without substantial difficulties. Nevertheless, it is possible

that the interviewed stakeholders are too optimistic on the possibilities for independent evaluations, as findings on the previous dimensions have demonstrated that e.g. the users of evaluations and their expectations are often not easy to identify. Therefore evaluations often take place 'within the system', as a result of which also external evaluations can be conceived rather narrowly, avoiding as well the more controversial issues. In these cases it is evident that there will be fewer problems within the institutional and political context, as a form of self-censorship has taken place.

4. Discussion and Conclusion

4.1. Main findings

Due to an increasing importance of evaluations within development cooperation, it has become all the more important to analyse if initial conditions for relevant and sound evaluations are met. As experiences with evaluations within Belgian development cooperation suggest that the initial conditions are often not met, the Office of the Special Evaluator commissioned an evaluability review with the aim to improve evaluability in the future. As a first step we elaborated a review framework which consists of three key dimensions, i.e. design of intervention and M&E system (theoretical evaluability), data availability and M&E system in practice (practical evaluability) and evaluation (organisational and broader) context. These key dimensions are further divided into eight components and 62 items. We used the review framework to analyse the evaluability of 40 interventions in 4 countries (Benin, DRC, Rwanda and Belgium). Drawing upon a combination of secondary and primary data collection, we scored each component and item on different OECD/DAC criteria (i.e. relevance, effectiveness, efficiency, impact and sustainability) using a five point grading scale.

The findings demonstrate that the evaluation context poses less challenges to evaluations than the quality of the intervention and M&E design and implementation of the M&E system in practice. This illustrates the margin for improvement of those dimensions that organisations have most control of. One weakness which can, for instance, easily be remedied for and which obviously affects evaluability is the translation of changes in an intervention into the M&E system, something which half of the eleven 'modified' interventions in our sample did not do.

Additionally, our findings highlight that flaws in the design often have consequences during implementation and thus affect evaluability directly and indirectly. This holds to varying degrees for 'monitoring' as compared to 'evaluation'. While initial (design) weaknesses can often be corrected for at the level of monitoring, they seem to affect more strongly the evaluation function. Deficient gender analysis or lack of attention for social differentiation within the target group at the outset of an intervention for instance risks to leave unexplored important underlying reasons for low performance and seriously undermines evaluative quality. Along the same lines, the strong focus on the lower (implementation) levels of the ToC at the moment of the design also risks evaluations to focus only on reality as it has been defined or interpreted by those involved in the intervention. This inward-looking focus is further aggravated by the limited attention at the outset of the intervention for the characteristics and (evaluation) interests of stakeholders beyond those immediately involved in the intervention, which also negatively affects the use/influence of evaluative findings.

Our findings about the lack of conditions in place for sound *impact* evaluation hint at the fact that the situation in Belgian development cooperation is as problematic as the one reported on by Wholey for the US about three decades ago (Wholey, 1979). However, broadening our scope beyond 'impact' to all OECD/DAC criteria allows to nuance the conclusion somewhat as the interventions under study generally score much better in terms of evaluability of *effectiveness* and *efficiency*. As mentioned above, in most cases this is already clear from the ToC, which is much better developed in terms of implementation theory as compared to impact theory.

The difference in evaluability among different OECD/DAC criteria does not necessarily come as a surprise. Organisations are much more incentivized to focus on the lower levels of the ToC as insights into those elements are directly important for the implementation of their own interventions and for adaptations that are directly applicable to these interventions. This is much less the case for higher levels in the ToC and related impact and sustainability evaluations which generally deliver findings and insights that are interesting beyond the own organisation's boundaries. This public goods character of impact and sustainability evaluations tends to incentivize free-rider behaviour whereby all organisations have an interest in waiting until other organisations invest in impact evaluations which finally leads to an underinvestment in impact evaluations (see also CGD, 2006). Adding to this is the fact that sustainability and impact evaluations are generally methodologically more difficult, expensive and politically more sensitive.

Against this background, it is somehow surprising that the main funding agency, the Directorate General for Development Cooperation, who is responsible for policy making and so by definition interested in evaluative findings beyond the implementation level, has, so far, not really incentivized implementing agencies to invest in satisfying some of the basic conditions for sustainability and impact evaluation. While there are some recent indications that things might be changing, the focus has to a large extent been on 'accountability' and the levels that interventions and implementing agencies have under their control.

Although most of the 40 interventions currently do not have the necessary conditions in place for sound impact evaluations, impact evaluations are almost 'ritually' being commissioned for almost all of these interventions which puts into perspective their quality and usefulness. From this vantage point, it might be important to invest first in the integration of evaluability assessment in interventions. While EA in itself will not put the conditions in place for sound evaluation, it allows to diagnose the degree to which different interventions are ready for different types of evaluations and

in which dimensions (design, practice or context) strengths and weaknesses are situated. As such EA is a useful instrument in the development and implementation of an overarching M&E policy.

In terms of evaluation policy, it might prove useful to develop a kind of flexible portfolio approach towards evaluation with different types of evaluations (with different evaluability conditions) being applied to different types of interventions. As Pritchett et al (2013) indicate, impact evaluations are specifically interesting and important for those interventions that have the possibility for upscaling. For types of interventions that have already demonstrated their impact in various settings, evaluations that focus on implementation are sufficient. An interesting route to explore is the practice of 'structured experiential learning' (MeE), introduced by Pritchett et al. (2013), which refers to evaluative activities ('e') between 'monitoring' (M) and 'evaluation' (E). Structural experiential learning is considered to be particularly useful for innovative interventions (with complex programme theories) where there is not enough knowledge yet regarding the effectiveness and efficiency of different implementation modalities. As discussed in Pritchett et al. (2013), in such circumstances a more flexible type of ex-ante design which underscores the 'unknown' and allows for experimentation with different implementation modalities and close monitoring and comparison of these different implementation modalities is advisable.

A same level of flexibility is warranted in terms of (impact) evaluation methodologies as not all interventions are apt for the same type of (impact) evaluation method. Similarly, a portfolio approach could also allow for some evaluations to be primarily driven by accountability objectives while others focus mainly on learning. In both cases, EA might be a useful instrument as different (impact) evaluation methods have different requirements in terms of design and data availability (see also footnotes 8 and 9) while also evaluations driven by different objectives need different elements in place, for instance, in terms of independence and degree of involvement of different key players (see context conditions).

Once a portfolio approach is getting shape, an incentive system needs to be put in place to motivate agencies and staff involved to invest in realizing the necessary conditions (in design, practice and context) for different types of evaluation. As regards impact evaluation for instance, it might be worthwhile to explore opportunities to set up a kind of pooled fund with different organisations to finance M&E expertise with each organisation having a number of expertise days available (kind of drawing rights). The recent initiative of DGD and a consortium of non-governmental actors who have pooled resources and jointly identified interventions for which an impact evaluation will be commissioned, is an interesting experiment in this regard. Such use of a pooled funding mechanism might counter the free-rider problem which is one of the underlying reasons for the underinvestment in impact evaluation.

While a pooled funding mechanism might counter the free-rider problem which currently leads to underinvestment in impact evaluation, it does not necessarily guarantee feedback of evaluative findings beyond the actors directly involved, nor their involvement in different phases of M&E processes. As also reflected in our findings, the neglect of local stakeholders as well as locally existing M&E practices is a much observed phenomenon with negative ‘evaluability’ implications, particularly in terms of an evaluation’s usefulness (and use) as well as M&E system sustainability beyond the intervention’s lifecycle. It is particularly also from this vantage point that it is important for EA to explicitly diagnose the degree of interaction and alignment with local M&E systems as well as getting a grasp of local contextual realities¹⁰.

4.2. Challenges and Limitations

Our own evaluability assessment was confronted with a number of challenges and limitations. First, while, as suggested in Strosberg and Wholey (1983), the purpose, process and limitations of the evaluability study had been communicated with the stakeholders involved, there was still

¹⁰ See WHO and IHP+ (2011) and authors (source deleted for blind review) for a more detailed discussion and suggestions on how to strengthen and align with national M&E systems and practices.

considerable resistance towards the study at the outset. As discussed above, this was to some extent related to confusions around the notion of 'evaluability' and the exact purpose of the evaluability study and mainly fuelled by suspicions on the potential consequences for interventions which obtained a poor evaluability score. This concern was not entirely unfounded as a study conducted by the Dutch Policy and Operations Evaluation Department (IOB) on the evaluability of development interventions resulted in the withdrawal of funding from organisations for which deficits in evaluability were identified. An additional explanatory note on the concept of 'evaluability' and the adjustment of the initial aim of the study towards a more exploratory and formative use of the EA attenuated the resistance while suspicions were never ruled out completely.

Second, the study team sometimes encountered challenges in identifying the required level of depth in the evaluability assessment. In other words, it was considered challenging to determine whether the mere availability of information was sufficient to assess the evaluability or whether the quality of information required a more in-depth analysis to come to a valid assessment of the evaluability. Along the same lines, the team sometimes struggled with delineating what should be included in an evaluability study without ending up doing an evaluation of the selected interventions, which is considered by Davies (2013) to be one of the principal risks of an evaluability assessment (see also Gilchrist, 2014). Finally, and as also indicated by D'Ostie et al. (2013) in their evaluability assessment of a West Africa based Non-Governmental Organisation, the threshold to determine if a condition is met or not was not always straightforward. The use of a checklist with a detailed scoring grid was particularly useful in this regard. While the joint elaboration of the checklist required a substantial time investment at the outset of the evaluability study, it allowed systematisation and transparency later on which also increased the validity and the usefulness of the exercise.

4.3. Where to next ?

As discussed in section 2.2, we have purposively constructed our sample as to allow comparative analysis and identification of factors that possibly associate with high or low evaluability. Some first crude analysis shows not much variation in evaluability among interventions in different countries or with different degrees of ToC complexity. What proves to be important is the implementing agency, with higher overall evaluability among interventions that are implemented by the BTC or NGOs as compared to those implemented by other actors such as cities, provinces, universities, etc. This does not really come as a surprise as development cooperation does not necessarily belong to the core mandate of the latter organisations which has implications in terms of the requirements set by DGD.

Interestingly, and somehow counter to expectations, interventions with a more complex ToC seem to score slightly higher than those with a lower level of complexity. Discussions with implementing agencies' staff, both in Brussels and on the ground, lend us to believe that more efforts are done in terms of analysis and design, both of the intervention and the related M&E systems, in cases where interventions are more complex and where donors still need to be convinced of their value. More in-depth analysis is needed to identify whether this holds for all different EA components and OECD/DAC criteria. Similarly, also the between-country analysis needs further refinement and cross-reading with previous studies on the politics of M&E. Particularly the findings for the Rwandan context do not match some of our own earlier findings (source deleted for blind review) which have demonstrated the influence a closed political opportunity structure might have on various dimensions of evaluations, such as the degree of independence, the evaluation focus, the level and layers of disaggregation, etc. Our current counter-intuitive findings could to some extent be related to a number of factors which generally affect evaluability and, likewise, may have compromised the quality of our own (evaluation) study. This includes among others social desirable answering, a (too) narrow demarcation of evaluation stakeholders, donors' confinement to 'easy' interventions (and evaluations) that do not address the more politically-sensitive governance issues. On the other hand, the Rwandan drive towards quick implementation, results-orientation and country-ownership might also put a number of building blocks in place which are conducive to evaluation. Obviously, more in-

depth between and within-case analysis is necessary to provide more insight into factors that lower or increase different dimensions of evaluability which may also feed into the formulation of more specific recommendations.

Acknowledgement

The authors would like to thank Ivo Hooghe (SEO), all interviewees and members of the evaluability study advisory board for their participation in the study.

Funding

The evaluability study was funded by the Belgian Special Evaluation Office (SEO).

References

- Bamberger, M., Rugh, J., Church, M., & Fort, L. (2004). Shoestring evaluation: designing impact evaluations under budget, time and data constraints. *American Journal of Evaluation*, 25 (1), 5-37.
- Caubergs, L., Charlier, S., Holvoet, N., Inberg, L., Van Esbroeck, D. (2014). *A difficult path towards equity: gender and development in Belgian Development Cooperation: Summary report*. Brussels: Special Evaluation Office of International Cooperation.
- Center for Global Development (2006). *When will we ever learn? Improving Lives through Impact Evaluation*. Washington, D.C: Center for Global Development.
- Dahler-Larsen, P. (2013). Evaluation as a situation or a universal good? Why evaluability assessments for evaluation systems is a good idea, what it might look in practice, and why it is not fashionable. *Scandinavian Journal of Public Administration*, 16 (3), 29-46.
- Davies, R. (2013). Planning evaluability assessments: A synthesis of literature with recommendations. *Working Paper 40*. London: Department for International Development.
- Davies, R., & Payne, L. (2015). Evaluability Assessments: Reflections on a review of the literature. *Evaluation*, 21 (2), 216-231.
- D'Ostie-Racine, L., Dagenais, C., & Ridde, V. (2013). An evaluability assessment of a West Africa based Non-Governmental Organization's (NGO) progressive evaluation strategy. *Evaluation and Program Planning*, 36, 71-79.
- Gilchrist, J. (2014). *Evaluability Assessment: Developing a Process to Determine if a Program is Ready and Able to be Evaluated* (unpublished master dissertation). Barnaby/Surrey/Vancouver: Simon Fraser University.

Inter-American Development Bank (2002). *Analysis of Project Evaluability Year 2001*. Washington DC: Inter-American Development Bank.

Inter-American Development Bank (2007). *Analysis of Project Evaluability Year 2005*. Washington DC: Inter-American Development Bank.

International Labour Organization (2012). *Dimensions of the Evaluability Instrument. I-eval Resource Kit, Guidance Note 12*. Genève: International Labour Organization - Evaluation Unit.

Leviton, L.C., Kettel, K. L., Rog, D., Dawkins, N., & Cotton, D. (2010). Evaluability Assessment to Improve Public Health Policies, Programs and Practices. *Annual Review of Public Health*, 31, 213-233.

Lloyd, R., Poate, D., & Villanger, E. (2013). Results measurements and evaluability: a comparative analysis. *Journal of Development Effectiveness*, 6 (4), 378-391.

McKinney, L. (2010). Evaluability Assessment: Laying the Foundation for Effective Evaluation of a Community College Retention Program. *Community Journal of Research and Practice*, 34 (4), 299-317.

OECD/DAC (2002). *Glossary of Key Terms in Evaluation and Results Based Management*. Paris: OECD/DAC.

Peersman, G., Guijt, I., & Pasanen, T. (2015). *Evaluability Assessment for Impact Evaluation. Guidance, checklists and decision support*. A Methods Lab publication. London: Overseas Development Institute.

Poate, D., Riddell, R., Chapman, N., & Curran, T (2000). *The Evaluability of Democracy and Human Rights Projects. A logframe-related assessment*. Stockholm: Swedish International Development Cooperation Agency.

Pritchett, L., Samji, S., & Hammer, L. (2013). It's all about MeE: Using structured experiential learning to crawl the design space. *Working Paper 322*. Washington D.C., Center for Global Development.

Rossi, P.H., Lipsey, M.W., & Freeman, H.E. (2004). *Evaluation: a systematic approach (7th edition)*. Thousand Oaks: Sage.

Schmidt, R., Beyna, L., & Haar, J. (1982). Evaluability Assessment: Principles and Practice. In G.J. Stahler, & W.R. Tash (Eds), *Innovative Approaches to Mental Health Evaluation*. New York: Academic Press.

Smith, M.F. (1990). Evaluability Assessment: Reflections on the Process. *Evaluation and Program Planning*, 13, 359-364.

Special Evaluation Office of the Belgian Development Cooperation (SEO)(2015). *Open offerteaanvraag voor de studie van de praktische valueerbaarheid van ge(co)financiered interventions van de Belgische Ontwikkelingssamenwerking*. Brussels: Federal Public Service Foreign Affairs, Foreign Trade and Development Cooperation.

Thurston, W., & Potvin, L. (2003). Evaluability Assessment: A Tool for Incorporating Evaluation in Social Change Programmes. *Evaluation*, 9 (4), 453-469.

Trevisan, M.S. (2007). Evaluability Assessment From 1986 to 2006. *American Journal of Evaluation*, 28 (3), 290-303.

Wholey, J.S. (1979) *Evaluation: Promise and Performance*. Washington DC: Urban Institute.

Wholey, J.S., Hatry, H.P., & Newcomer, K.E. (2010). *Handbook of Practical Program Evaluation*. New York: John Wiley & Sons.

World Health Organization (WHO) and International Health Partnership (IHP)+. (2011). *Monitoring, Evaluation and Review of National Health Strategies: A country-led platform for information and accountability*. Geneva: WHO & IHP+.

Appendix A. Review framework, including scores for the dimensions, components and items

	relevance	effectiveness	efficiency	Impact	sustainability
1. Analysis of the intervention design (theoretical evaluability)					
1.1 The underlying analysis	66.25%	65%			
1.1.1 The target groups are clearly identified (demarcated) and described	62%	62.5%			
1.1.2 The rationale of the intervention and the (problem) situation of the target group(s) is clearly described	71.25%				
1.1.3 The role of the target group(s) is clearly described	65%	63.75%			
1.1.4 The role of the major stakeholders (exclusive of target group) is clearly described	63.25%	61.75%			
1.1.5 Gender analysis is integrated in the analysis	35%	35%			
1.1.6 The link between analysis and intervention objectives is clearly described	75%				
1.1.7 The link between the rationale of the intervention and the sector policy of the partner country is clearly described	77.5%				
1.2 The intervention logic and the theory of change (TOC)		50%	62.5%	27.5%	56.25%
1.2.1 A clear and correct distinction is made between outputs, outcomes and impact		66.25%		41.25%	
1.2.2 The ToC from outputs to outcomes and impacts is clearly elaborated		58.75%		22.5%	
1.2.3 The ToC is logic and realistic		78.75%	85%	55%	
1.2.4 Critical and crucial links/ingredients in the ToC are identified and can be tested		51.25%	58.75%	21.75%	
1.2.5 The ToC includes concrete measures (input, activity and output level) to safeguard the sustainability of the intervention results					62.5%
1.2.6 The internal risks are clearly identified and explored/estimated		50%	52.5%		45%
1.2.7 The external assumptions are clearly identified and explored/estimated	45%	52.5%	55%	28.75%	47.5 %
1.2.8 The allocation of the resources foreseen (investments, personnel and operational costs, inputs from other stakeholders) to the outputs is clear			66%		
1.3 The proposed M&E system	36.25%	45.75%	46.25%	23.75%	36.5%
1.3.1 The most important results envisaged are made operational in an adequate way		53.75%	66.25%	20.75%	43%
1.3.2 Where necessary/relevant indicators are disaggregated by sex or other relevant characteristics		20.5%	20.25%	8.5%	15.75%
1.3.3 The proposed M&E system includes a consistent translation of the proposed intervention logic and ToC	50%	61.25%	66.25%	32.5%	42.25%
1.3.4 The method for monitoring and evaluating the intervention results and their sustainability is clearly described	42.5%	53.75%	57%	28.25%	40.5%
1.3.5 The method to monitor the assumptions is clearly described	21.25%	21.25%	22.5%	13.75%	20.5 %
1.3.6 The method to monitor the internal risks is clearly described	17.5%	20%	21.25%	11.25%	17 %
1.3.7 The human and financial resources involved in the M&E system are clearly described			35%		
1.3.8 The proposed MIS allows an allocation of expenditures to specific outputs and intervention components			66.75%		
1.3.9 The way in which the M&E system of the intervention is aligned/related to the national/local M&E system is clearly described	14.25%	17.25%	15.75%	14.25%	14.25%

	relevance	effectiveness	efficiency	Impact	sustainability
1.4 Consistency and adaptation of the intervention logic and ToC	63.75%	61%	63%	62.5%	68.75%
1.4.1 Changes in the intervention logic and the ToC are clearly reported and explained	81.75%	75%	69%	80%	79.25%
1.4.2 Information is available about the vision and the opinions of the most important stakeholders regarding changes in the intervention logic and ToC	45.5%	40.75%	35.75%	45%	41.75%
1.4.3 Changes in the intervention logic and ToC are adequately translated and integrated in the M&E system	54.5%	56.25%	52.5%	50%	58.25%
2. Practice regarding intervention implementation, intervention management and context (practical evaluability)					
2.1 Availability of basic information about intervention implementation	48.75%	55%	63.25%	28.25%	40.5%
2.1.1 The basic documents (that could reasonably be expected) are available	75.75%	76.25%	78.25%	68.25%	73.75%
2.1.2 Baseline information (consistent with the intervention logic) about the target group(s) is available (for relevant indicators)		62.5%	72.5%	36.25%	46%
2.1.3 Baseline information (consistent with the intervention logic) about the counterfactual is available (for relevant indicators)		7%	7%	11%	7.25%
2.1.4 Disaggregated baseline information according to gender (or other relevant parameters) is available on relevant indicators		19%	21.5%	12.75%	
2.1.5 Information (that could reasonably be expected) about the progress of the intervention implementation is available		59.25%	73.25%	17.25%	
2.1.6 Information about the participation of the initial target group is available	56.5%	57.75%	60.25%	46%	52.75%
2.1.7 Information is available about the data collection process of the indicators	43.25%	52%	57%	32%	42.25%
2.1.8 The data collection proposal allows in principle reliable data collection of the indicators	50%	53%	61.75%	38%	36.25%
2.1.9 Information about the monitoring of internal risks is available and potential consequences for the intervention logic and intervention implementation are indicated		35%	43.75%	28.25%	46.5%
2.1.10 Information about the monitoring of (external) assumptions is available and potential consequences for the intervention logic and intervention implementation are indicated	30.25%	34.25%	40%	30%	46.25%
2.1.11 Intervention expenditures are well documented/ recorded and linked to the outputs			73.75%		
2.2 The M&E system (in practice)	50.75%	57.5%	62.5%	38.25%	46.25%
2.2.1 The vision on and the role of monitoring and evaluation (both M&E and independent evaluation) are clear		55.75%	57.75%	47.5%	43.25%
2.2.2 The most important stakeholders agreed upon the proposed M&E system and their role therein	79.5%	80%	83.25%	80.75%	79.5%
2.2.3 There are sufficient resources (time, financial resources, human resources) provided to allow for an adequate functioning of the M&E system.	66.75%	71.75%	78.75%	55.25%	65.75%
2.2.4 The responsibilities and procedures for collecting and analyzing M&E data have clearly defined	62.5%	71.25%	77%	41%	60.25%
2.2.5 The responsibilities and procedures for decision-making based on the analysis of M&E information are clearly defined	65.5%	93.75%	75.75%	48.75%	64%
2.2.6 The staff responsible for M&E is competent and independent	48.75%	52.75%	58%	43.25%	50%
2.2.7 The M&E system of the intervention is (as much as possible) related/aligned to the national/local M&E system	39.75%	47%	53%	38%	41%

	relevance	effectiveness	efficiency	Impact	sustainability
2.2.8 There is an internal drive for strategic management and learning	68%	70%	72.5%	67.5%	68%
2.2.9 The M&E results are being used for learning	51.25%	52%	57.5%	48.25%	50%
2.2.10 The M&E results are being used for accountability	55.75%	60%	65.75%	48.75%	52%
2.2.11 Evaluations/studies have been conducted which are of good quality and these provide useful information (relative to what could have been expected)	41.75%	44.5%	43.75%	30.5%	39.5%
2.2.12 The M&E system is regularly reviewed on its quality and adjusted accordingly	43.5%	47.5%	50%	45%	46.25%
3. The evaluation context					
3.1 Attitude of the key players	67%	67.5%	67.5%	64.5%	65.75%
3.1.1 The principal users of the evaluation and their expectations/interests with respect to the evaluation are clearly defined	34%	34%	34%	31.75%	33.25%
3.1.2 The key stakeholders express demand for (or are at least interested in) evaluation	61.25%	61.25%	61.25%	60%	61.25%
3.1.3 The expectations of the key stakeholders with respect to evaluation are mutually compatible	77.75%	77.75%	79.75%	77.75%	79%
3.1.4 The expectations of the key stakeholders about evaluation are realistic (in relation to the available resources)	83%	80.75%	87.25%	77.75%	83%
3.1.5 The principal users were/will be involved in the evaluation process.	67.5%	66.75%	66.75%	68.5%	68.5%
3.1.6 The most important stakeholders (inclusive of the beneficiaries) were involved/will be involved in the evaluation process	63.25%	63%	64.5%	63.75%	63.75%
3.1.7 The relationships amongst the key stakeholders are “healthy”	81.5%	81%	81%	80.5%	80.5%
3.1.8 It is possible to contact all stakeholders without risking reciprocal influence	90.75%	91%	91%	90.75%	90.75%
3.1.9 There is a positive attitude towards independent evaluation amongst all key stakeholders	79.5%	79.5%	79.5%	80.25%	79.5%
3.2 The broader context	77.5%	81.25%	81.25%	77.5%	78.75%
3.2.1 The broader institutional and political context is positive towards independent evaluation.	85%	86.25%	86.25%	85%	86.25%
3.2.2 The socio-cultural context at the level of the target group(s) allows for the adequate collection of information	83.75%	86.25%	86.25%	83.75%	85%
3.2.3 Local expertise with the required profile for evaluation is available	81.25%	82.5%	81.25%	80%	81.25%