Reflection Paper

Huib Huyse (HIVA-KU Leuven)
Wenny Ho (Hivos)
Karel Chambille (Hivos)

LEARNING AND ACCOUNTABILITY IN TURBULENT WATERS

Reflection on the Evaluation of the Citizen Agency Consortium Advocacy Programs

KU LEUVEN HIVA
RESEARCH INSTITUTE FOR WORK AND SOCIETY

Hivos people unlimited

EVALUATION OF ADVOCACY PROGRAMS: LEARNING AND ACCOUNTABILITY IN TURBULENT WATERS

TABLE OF CONTENTS

1. **The original evaluation framework and process** 07
2. **Reflection on the main changes in the evaluation designs and implementation** 11
3. **Did the evaluation approach deliver? Five critical issues** 14
   - 3.1. **Did monitoring contribute to evaluation?** 14
   - 3.2. **Achieving depth and breadth in evaluating a multi-country/ multi-thematic advocacy program** 17
   - 3.3. **Evaluation for whom?** 19
   - 3.4. **Evaluating in times of corona** 22
4. **Discussion** 24

INTRODUCTION

With Dialogue and Dissent, the main funding track for Dutch development civil society organizations (CSOs) in the 2016-2020 period, the Dutch government chartered into unknown territories. Not only because of its strict focus on lobby and advocacy and the specific implementation modalities, but also on monitoring and evaluation (M&E). This methodological note describes the experiences of a consortium of international development CSOs with implementing a large multi-country evaluation of four programs, funded through this unusual funding framework. In this note, we take stock of what worked and what didn't work when evaluating the outcomes of a diverse set of lobby and advocacy strategies and activities, and assessing the effects of the capacity development efforts. We outline a number of lessons for future evaluations of lobby and advocacy. In a final 'Discussion' section, we also reflect on what these findings could imply for future monitoring and evaluation efforts of advocacy programs – in terms of defining what counts as outcomes, how to do contribution analysis, and how to support learning.

Dialogue and Dissent (D&D) was a drastic departure from the past, and remains quite unique compared to CSO funding frameworks of other OECD-DAC donors. Its exclusive focus on lobby and advocacy, and the requirement for each D&D program to develop partnerships with the Dutch government, both at the policy and operational levels, stood out. Finally, D&D wanted to move away from managerial project management approaches that push CSOs towards long-term detailed planning, predictability and control. Complexity was embraced through lighter, process-oriented and more flexible procedures, for example through the use of theories of change. The main policy note (Kamstra, 2017) underpinning D&D had concluded that CSO-government relationships had become too managerial under the previous funding line (MFS-2), hindering the transformational potential of CSO programs.

This methodological reflection note draws lessons from the final evaluation of the Strategic Partnership Citizen Agency Consortium (SP CAC) program, implemented by Hivos, a Dutch NGO, the International Institute for Environment and Development (IIED) and Article 19. The SP CAC was funded through the Dialogue and Dissent framework (2016-2020) and includes four distinct lobby and advocacy (L&A) programs **(see page 5).**

Under D&D, monitoring and evaluation guidelines from the ministry were less prescriptive than was the case in earlier funding tracks. Hivos used the flexibility to strike a user-oriented balance between accountability and learning. This translated into an evaluation design comprising: a set of in-depth case studies based on learning topics that program staff identified, and a program-wide assessment of reported outcomes from outcome harvesting, as well as other monitoring and evaluation sources. After an open, competitive bidding process, Hivos seclected four consultancy agencies to each carry out one of the four program evaluations.

The evaluation process was complex, given that it covered multiple L&A programs, with different actors working on rather sensitive topics in a wide range of low- and middle-income countries. An external evaluation reference group was established in 2017 to provide methodological orientation prior and throughout the evaluation process. The writing of this reflection note is the result of a collaboration between Huib Huyse (HIVA-KU Leuven), who was part of the evaluation external reference group (ERG), and Wenny Ho and Karel Chambille who were Hivos' overall evaluation managers of the SP CAC program

Citizen Agency program

The **Decent Work 4 Women** (DW4W) program focused on fair wages, safety and security in the workplace, good working conditions, in particular targeting women, in the horticulture sector in Eastern and Southern Africa, and the Netherlands.

Green and Inclusive Growth (GIE) sought to meet people's energy needs through green and inclusive energy systems that create economic opportunities for women and men while mitigating climate change (Central America, Nepal, Kenya, Tanzania, Zimbabwe, Malawi, Guatemala, Myanmar, and The Netherlands).

The **Sustainable Diets for All** (SD4ALL) program aimed to make more sustainable, diverse, healthy, and nutritious food available to low-income citizens (Bolivia, Zambia, Uganda, Kenya, Indonesia, and The Netherlands).

Open Up Contracting (OC) aimed to give people equal access to quality public goods and services through building the capacity of infomediary partners, brokering relationships and using an evidence-based combination of lobbying and advocacy to enhance transparency, participation, and accountability. (Bolivia, Guatemala, Kenya, Malawi, Tanzania, Indonesia, Philippines, and the Netherlands).

© Makmende Media/Hivos

01

THE ORIGINAL EVALUATION FRAMEWORK AND PROCESS

The evaluation framework that the Hivos evaluation managers designed combined a theory-based evaluation logic with 4-5 case studies per program. The evaluation teams were expected to build on existing monitoring efforts. Over time, all the programs had systematically collected monitoring data on lobby and advocacy outcomes according to the Outcome Harvesting (OH) methodology. These datasets served as a basis for both program-wide analysis and case study work. In addition, two out of four programs documented stories from advocates on specific L&A trajectories through the Narrative Assessment[1] (NA) methodology. NA takes the day-to-day experiences and strategic reflections of the advocates as an important source of information. The monitoring of capacity development differed between the four programs.

In late 2018-early 2019, first conversations were held for the external end evaluation. To maximize the learning potential and usefulness for the programs, the four global program managers were asked to propose possible learning topics for case studies. Some of the learning topics emerged from the internal mid-term reflection exercise within and across the four CAC programs.[2]

The overall Terms of Reference (TOR) provided for a four-phase evaluation process:[3]

1. A substantiation of the portfolio of outcomes, harvested by the program since the beginning of 2017, to increase the credibility of the monitoring data, as an input for the evaluation teams;
2. Four parallel thematic evaluations, one of each thematic program;
3. A comparative study of the CAC organizational & partnership aspects that might have influenced the program, based on the four thematic reports;
4. A synthesis exercise bringing together the findings from the different studies on topics that emerged across all sub-evaluations for its relevance to current and future programming and implementation.

The evaluation questions in the ToR provided an operationalization of the evaluation criteria's effectiveness, relevance, sustainability and efficiency. The ToR further specified that the thematic evaluations should balance an overall analysis of each thematic program with a number of in-depth case studies on the selected learning topics. Starting from a description of changes – in agendas, policies and practices of targeted social actors and in the L&A capacities of participating organizations – the thematic evaluations were asked to compare these changes with the program objectives. They were also asked to assess inclusiveness and potential effects on climate change, as well as the relevance of these changes in the context of the program countries. Furthermore, the evaluations assessed the contribution of the programs to the observed changes, the sustainability of these changes, and finally made an attempt to analyze aspects of efficiency.

THE EVALUATION PROCESS

The OH substantiation exercise was commissioned to three independent evaluators with OH expertise. They had not been involved in the OH monitoring.[4] The four program managers were asked to select a purposive sample of harvested outcomes, based on their importance for the program and/or need for validation.

Four evaluation teams were finally contracted in November-December, 2019. In their inception reports, they (further) elaborated their proposed approaches towards answering the evaluation questions and their sample selections of country/case studies aligned with a program's characteristics and interests.

1. More information on https://www.hivos.org/news/narrative-assessment-bringing-out-the-story-of-your-advocacy/
2. The evaluation managers and the CAC program managers felt at the time that a mid-term reflection process would draw out more opportunities for learning than a traditional evaluation would do.
3. This paper focuses mostly on the evaluation activities under 1 and 2.
4. In 2017, an earlier and more limited substantiation exercise was commissioned to the two OH experts who accompanied the OH monitoring process.

sources of information, combining the study of documents with interviews, focus group discussion (FGDs) and participant observation. In the case of the GIE program, narrative assessments were conducted to deepen aspects of the case studies. Program monitoring data served as inputs for the evaluation teams, especially the harvested (substantiated) outcomes. The evaluation teams applied specific causal analysis on a sample of these harvested outcomes. For their analyses of capacity development, the evaluators depended more on their own data collection, as monitoring data was less complete. Efficiency was assessed through the Efficiency Lab approach.

The thematic evaluations produced 17 country/case study reports and four overall thematic reports. The draft country/case study reports were presented to staff and partners for comments and validation. The key thematic program staff, the two evaluation managers and ERG then assessed the overall thematic reports.

The original evaluation plan had included the hosting of thematic learning events with staff and partners and the evaluators in the various regions. Due to COVID-19, these plans had to be changed and a number of online learning events were organized instead.

Phase three of the evaluation was a comparative secondary study of the four thematic evaluation reports, and other documents of relevance[5] to the evaluation object.

The objective of the final synthesis exercise was to bring together findings from the four thematic programs around topics that are considered relevant beyond the Strategic Partnership Citizen Agency Consortium.

5. Hivos policy documents; policy documents from the Ministry of Foreign Affairs and the Ministry's Policy and Operations Evaluation department (IOB; relevant study reports.)

© Sven Torfinn/ENERGIA

02

REFLECTION ON THE **MAIN CHANGES IN THE EVALUATION DESIGNS AND IMPLEMENTATION**

This section describes how the different thematic evaluation teams translated the original Terms of Reference (ToR), including the evaluation questions, into a methodological and operationalized approach, and the results of that. Factors at play were the nature of the programs under evaluation, methodological preferences, and/or external circumstances.

Methodological uniformity and diversity – Overall, the four evaluations kept to the original methodological design, although each complemented it with different methods and frameworks.

The degree of elaboration of the *evaluation frameworks* underpinning each evaluation differed. While the GIE evaluation team used a simple evaluation framework with evaluation questions and sub-questions, the SD4ALL and DW4W evaluation teams further added judgement criteria. These criteria aimed at improving the consistency and transparency between the different evaluators, ensuring that they consider the same criteria when judging an evaluation item. A similar reasoning led to the introduction of rubrics for each indicator by the OC evaluation team. Beyond improving consistency in the criteria used, rubrics more explicitly define the conditions under which the evaluators should score an evaluation indicator as very poor, poor, good or very good.

Regarding the *program-wide assessment,* all the evaluation teams used the outcomes harvested in the monitoring cycles to get a sense of what the program had achieved in different areas. The evaluation teams also used the opportunity of fieldwork at the country level to gain an understanding of the effects of the program, beyond the specific case studies in that country. Adopting a program's ecosystem perspective (for the OC evaluation) or the nexus approach (for the GIE evaluation) also helped to assess the strengthening of a system of partner organizations at the country and regional levels. All the evaluation reports state that the case studies were not selected for their representativeness, but as learning-oriented, information-rich illustrations, which means that the findings could not be used to generalize about the program. However, for one program (SD4ALL) the case studies covered data collection in almost all the partner countries (four out of five), implying that the case study findings covered a substantial part of the overall program. All in all, while some methodological differences could be observed, the overall approach of this part of the evaluation was the same in the four evaluations. However, some evaluations further elaborated criteria, to ensure consistency in assessment and strengthening transparency towards the program team.

While all evaluations used the OH monitoring data as the backbone for the *case studies,* there were some differences in how the case studies were selected and in the use of complementary monitoring data. The selection of case studies was informed by a set of learning questions as well as a screening of the available harvested outcomes. The learning questions were identified by the program teams at different points of time: some were identified during the mid-term reflection exercise, some during internal program meetings, while others were added or refined in preparation of the final evaluation. In the run-up to the evaluation, program teams were also asked to make suggestions for the case studies. For all the program teams, the selection of case studies was

mainly inspired by the learning questions. The evaluation teams considered this pre-evaluation exercise, but some made substantial changes to the shortlist based on practical and feasibility criteria. More specifically, the DW4W evaluation used an extensive analysis of the reported outcomes to shortlist a set of substantiated outcomes (4-5 outcomes per country) which could be used as cases rather than the earlier broadly defined set of cases based on learning questions. In a follow-up step, the shortlisted cases were then checked if they covered the learning questions sufficiently, including together with additional criteria (coverage of the different outcome domains, the different levels of change, the mix of partners, etc.). For the GIE and SD4ALL evaluations, the evaluators kept to the suggestions from the program teams because they were found to fit well with the overall and country-specific evaluation and learning questions. For the OC evaluation, however, each case study was tied to a specific theme, which was then studied across the whole set of program operations. This resulted in a more program-wide analysis for each case study compared to the other programs where case studies were linked to one country or a region. To summarize, some differences could be observed in how the case studies were finally selected, but substantial time was spent in selecting the cases for each evaluation. Each evaluation team considered the inputs from the program managers but critically reviewed the suggestions based on methodological and practical constraints.

Two program teams had collected monitoring data on the lobby and advocacy interventions through the NA approach, which was also made available to the evaluation teams. In addition, for the GIE program, the evaluation team also used in-depth interviews[6] based on the NA approach to further document specific L&A activities and obtain a better understanding of how they were achieved.

The SD4ALL evaluation used multi-stakeholder coalition theory to assess the different coalitions and multi-stakeholder initiatives against the seven principles for effective multi-stakeholder collaborations[7] The framework has mainly been used to define specific judgement criteria in the evaluation framework for this part of the SD4ALL program.

In addition, different approaches[8] to measure the efficiency of the program were used. Two evaluations (SD4ALL and DW4W) used a theory of efficiency approach to measure program efficiency through a light-touch version of the Multi-Attribute Decision Making (MADM) method. This method allows "program stakeholders to assess the 'usefulness' of a number of interventions in realizing program outcomes (from the ToC) against the amount of resources (time, money, effort, energy) needed to realize said outcomes (Synthesis report final evaluation SD4ALL, 2020, p59)."

The next section describes the similarities and differences in the approach used by the four evaluation teams.

Contracting four different evaluation teams helped ensure that the thematic expertise of the teams would fit the four programs. Although there was a risk that this could potentially lead to a lack of coherence in the evaluation methodologies, this did not occur. While the four teams did adapt the proposed overall methodology to the local context of each program and their methodological preferences, all maintained the combination of a program-wide assessment with case studies. Furthermore, the used evaluation methodologies did not diverge to the extent that it hampered comparison across the four programs. Allowing for differences in methodological approaches, while maintaining quality oversight, especially during the inception phase, did bring with it the advantage that evaluation teams were comfortable with the methodologies they used.

Missing or incomplete monitoring data – A major challenge across the four evaluations were the gaps and divergences in monitoring approach and data for capacity development (see also section 4.1). Before the launch of a guidance paper on the 5C capacity development model in 2016, the DW4W program developed a self-assessment form, which some CAC programs used, but not consistently. Others used their own approach or frameworks that external consultants had developed. After some initial test-runs during the first monitoring cycle, most program teams stopped using the 5C self-assessment forms (see section 4 on how monitoring data contributed to the evaluation) or replaced it with a simplified and/or informal approach.

6. All NA interviews had to be done online due to Covid-19 which hampered the quality of the interviews.
7. Brouwer H. and Woodhill J. with Hemmati M., Verhoosel K. and van Vugt S. (2016) The MSP Guide. How to design and facilitate multi-stakeholder partnerships. Wageningen University & Research, CDI The Netherlands. Retrieved from http://www.msp-guide.org/sites/default/files/case/msp_tool_guide.pdf.
8. Theory of Efficiency approach.

As this aspect covered one of the two main evaluation questions in the ToR, the evaluation teams had to reflect on how to deal with this gap in the inception phase. Two evaluation teams (SD4ALL and DW4W) combined a survey approach on capacity development for the whole program with interviews and focus groups, and the GIE evaluation team simplified the 5C framework to assess the observed capacity changes. The OC evaluation team did not use the existing capacity assessments but rather replaced the data gaps with interviews. This issue remained, however, a weaker point in the overall data collection process across the programs (see also section 3), because documenting capacity changes remained a weak spot and there was limited monitoring data to work with. This added to the evaluation teams' tasks to reconstruct what had happened.

A second challenge related to monitoring data from outcome harvesting. The programs had harvested outcomes until June 2019. However, with data collection activities running until March-April 2020, nine to ten months of the operations were not covered by OH monitoring data. While all the evaluation teams had the option of collecting additional outcome statements for this period, it took some time before this issue was on the radar. Only one evaluation team (DW4W) conducted a specific data collection activity to collect additional outcomes. Other evaluators used interviews, focus groups and document review to fill in some of the gaps. The OC program team ended up doing a round of OH monitoring for the missing period, but only managed to provide the dataset to the evaluation team at a late stage. Including these additional outcomes delayed the completion of the OC evaluation.

Changing theoretical concepts guiding the CAC program – Citizen Agency is a central concept in the overall narrative about the CAC SP program but it failed to emerge clearly as a concept-in-use during the evaluation inception phase. While certain aspects of how Hivos defines citizen agency did feature implicitly in the four programs, the evaluation teams could not identify a meaningful and uniform way of assessing whether the concept had figured in the reported outcomes. It therefore does not appear extensively in any of the evaluations. On the other hand, several evaluation teams observed the (implicit) attention for ecosystem thinking across the programs, with implications for capacity development (to be understood beyond the capacity of individual organizations), partner selection (to strengthen an ecosystem), and sustainability (to achieve a resilient support system that can continue to work on the topic after completion of the program). The OC evaluation team used ecosystem concepts as a central framework for the evaluation through specific sub-questions addressed in the report. The GIE evaluation made extensive use of the nexus framework, adopted by the GIE program team, to assess the specific articulation of the GIE's ecosystem concept. As the program aimed to achieve outcomes around the role that renewable energy could play in specific domains (e.g. gender, education or health), the thinking around this nexus then guided the selection of nexus partners (additional partners that could build linkages with key players on gender/education/health) and nexus targets (e.g. Ministries of Women's Affairs).

03

DID THE EVALUATION APPROACH DELIVER? **FIVE CRITICAL ISSUES**

3.1. DID MONITORING CONTRIBUTE TO EVALUATION?

As described in the previous sections, the two main monitoring datasets accessible for the evaluation teams were the outcome harvesting data, capturing the effects of various policy influencing efforts, and the organizational/networking capacity assessments (OCA). The perceived usefulness of both datasets differed substantially.

Outcome harvesting: 'Good foundations to build on, but further tweaking required'

The introduction of outcome harvesting in the 2016-2020 CAC program required substantial time and resources from both Hivos and its partners, but formed a part of the capacity development of the SP CAC. Collecting outcome statements was new for all partners, and most staff members. Compared to the existing monitoring practice based on logframes, it required a more holistic reflection on L&A outcomes during each monitoring cycle, and a critical review of the specific relevance for and contribution of the program to the observed changes. In preparation of the evaluation, Hivos had asked external consultants to complement the earlier substantiation exercise. The substantiation process involved contacting three external respondents for each outcome, asking them to confirm, refute and/or complement the outcome.

The two substantiation exercises together (2017 and 2019) covered 186 outcomes (i.e. 44% of the total outcomes harvested until mid-2019). Although time-consuming in the run-up to the evaluation, the evaluation teams assessed it positively as it allowed them to focus their efforts on the further validation and assessment of the reported contribution stories. At the same time, there was some unclarity about the status of outcomes that had received one or two positive confirmations from external stakeholders and one or two contested responses. It is unclear how the evaluation teams dealt with cases of divergent responses.

Some further insights emerged from using this pre-validated monitoring dataset in a formal evaluation process: Overall, the experiences are rather positive across the four evaluations. The reported outcome statements provided a solid basis to get a first understanding of the reach, scale and context of a program in the inception phase of the evaluation. A meta-analysis of all the statements, for example, helped to detect broad trends and allowed comparing the outcomes versus original goals. It also helped to identify outcomes that required further validation in the case studies. Outcome harvesting in combination with other tools and approaches such as theories of change, also introduced some level of evaluative thinking to the program teams before the evaluation had actually started, in ways that are feasible and useful for lobby and advocacy teams.

Using outcome harvesting monitoring data for evaluation could, however, be further improved and systematized. First, several evaluators indicated that a monitoring system based on outcome harvesting does not take away the need for output-level data. While outcome statements provide basic information about the chain of events in lobby and advocacy activities, this is often too sketchy and limited to have a deep understanding of the process of change leading from inputs to the reported outcomes.[9] This is an inherent limitation of the OH method. Two program teams opted to use Narrative Assessment (NA) to complement the output level data obtained through

9. This was one of the central reasons to develop the Narrative Assessment method that was used in two of the four programs.

OH. The GIE program team used NA monitoring data to articulate their way of doing lobby and advocacy in a learning document. The NA stories were also provided to the evaluation team. In the case of the SD4ALL program, the evaluation team was provided NA monitoring data that was already used as an input into the midterm reflection exercise. The SD4ALL evaluation report documented that NA supported the program team to gain a deeper understanding of structural drivers underpinning the change process, and as a source of information, contributed to annual reflection and planning processes.

Along similar lines, where relevant and feasible, other types of outcome-level data should still be collected, for example through research projects, and secondary sources of information.

Second, it is worth investing time and resources in processes that support the formulation of good quality outcome statements as well as in collective reflection on those outcomes. Some lessons can be drawn from existing practices in SP CAC. The quality of the monitoring and learning process was strengthened in the (sub-) programs that integrated additional training on OH, and/or had built collective data collection and sense-making activities into the project cycle to review the ToC against the reported outcomes. Writeshops and annual learning-oriented sense-making workshops were identified as good instruments to achieve this in a systematic way. Third, there was discussion in the evaluation teams about how to deal with outcomes that had not been substantiated versus those that had been. All the reported outcomes were considered in the evaluations, but for the case studies and specific causal analysis, mainly substantiated outcomes were incorporated.

Fourth, the outcome harvesting methodology guides evaluators in gaining a basic indication of the likely contribution of a program through the reported contribution statements and the external substantiation exercise, but this is insufficient when stronger indications of causality need to be established. In the case of lobby and advocacy programs, this requires an assessment of rival explanations, a rich description of the change process, and a broader consultation of key informants. This requires additional data collection steps and a systematic and transparent


© Sven Torfinn

presentation of the evidence. The extent to which this happened differed between the four evaluation teams. One evaluation team (GIE) used traditional triangulation techniques to assess the achievement of progress and contribution. A more elaborated and explicit approach was used in the other evaluations. The OC evaluation used a light version of contribution analysis in which contribution stories were constructed and assessed with the use of extensive referencing to primary and secondary sources, an assessment of the quality of the evidence, and rubrics to improve transparency in the assessment. The SD4ALL and DW4W evaluations did not use rubrics but fine-tuned the contribution analysis with contribution tables and the assessment of rival explanations.

Finally, as described in section 2, getting the timing right is essential. Documenting outcomes soon after an advocacy episode reduces the chance that relevant insights are lost and adjusting monitoring cycles to evaluation exercises guarantees that the evaluation teams have access to up to date outcome harvesting datasets.

Organizational/networking capacity assessment instruments: 'Useful by exception'

The experiences with the monitoring data from the organizational/networking capacity assessment (OCA) instruments were quite different. Over time, the assessment of the capacity of partners and networks became rather fragmented across two levels of Hivos' program cycle: one is a Hivos wide obligatory checklist when new partners are being contracted, and the other concerns the instruments used for the assessment, planning and monitoring of capacity development efforts in the SP CAC programs. The evaluation only worked with the latter.[10] As described in previous sections, the assessment, planning, and monitoring of capacity development in the CAC programs was a decentralized activity, largely left to the program teams in the countries. The 5 capability framework (5C) has a long history inside the development sector. It was introduced in response to earlier OCA frameworks which were found to be too reductionist and managerial, focused mainly on hard management skills (planning, accounting, project management,) and agnostic to the emergent nature of capacity. While the 5C framework has merits in bringing a systemic perspective to organizational change, the generic descriptions of the five capabilities tends to alienate practitioners. Hivos was aware of some of the critiques on the use of the 5C framework for monitoring and had produced a guidance note on the use of 5C (Hivos 2016), followed with centralized revised guidelines in 2017. Support for its implementation was organized per program. A result of the decentralized way of working and a lack of centralized steering within Hivos to support the monitoring of capacity development strategies – as was the case for OH – was that programs and countries used different, sometimes undocumented, approaches. DW4W developed a survey-style assessment tool that was then adopted by some other programs. Others used other approaches inspired by the 5C framework. Most program teams stopped reporting with these 5C instruments after the first monitoring round because the 5C framework was still found to be too conceptual, complex and not user-friendly. A second critique related to the incapability of the framework to capture the day-to-day reality of the extremely diverse set of Hivos partners. Third, in most cases the 5C assessments did not provide sufficient support for the drafting of capacity development plans as it lacked pointers to decide on the way forward. Finally, evaluators indicated that it was difficult to link the findings of the available 5C assessments, which varied in quality, to policy influencing outcomes.

The evaluation reports show at least two examples of positive experiences with OCA frameworks inside the evaluation process. In the SD4ALL and DW4W evaluations, the 5C framework was not used for data collection but rather for a meta-analysis of the capacity results. In the GIE evaluation, the evaluation team constructed a simplified OCA framework with 3 key areas of change, deducted from first insights in the inception phase and integrated with the overall capacity areas brought forward by the Ministry of Foreign Affairs.

To conclude, while the outcome harvesting monitoring dataset was found to be largely useful as an input into the evaluation process, although with some hiccups and limits, the story was different for the capacity development monitoring data. Although informal monitoring practices filled in some of the gaps emerging from the lack of follow-up and follow through at the program and country levels, most evaluation teams were confronted with incomplete datasets in this area.

10. The evaluation teams might not have been aware of the OCA exercises done in the partner selection phase.

3.2. ACHIEVING DEPTH AND BREADTH IN EVALUATING A MULTI-COUNTRY/MULTI-THEMATIC ADVOCACY PROGRAM

In the ToR, an evaluation design was outlined to provide in-depth insights through a limited number of case studies around predefined learning topics, as well as an overall assessment of each of the four programs through cross-portfolio data collection activities and the review of secondary data. As described above, during the inception phase, this generic framework was later (slightly) adapted to the thematic focus and context of each program as well as professional preferences of each evaluation team. This section reviews what can be said about how the four evaluation processes and reports dealt with the depth and breadth of ambition in their analysis: what worked well, what were some of the limitations and lessons learned?

Do the case studies provide a rich and contextualized analysis?

While three out of four evaluation teams had to organize a substantial part of their data collection via online interviews due to COVID-19, the country reports of all four evaluations build convincing and solid arguments for the different cases. Where data was incomplete or sporadic, this is mentioned explicitly in the analysis. The OC report makes the quality of the evidence more transparent by using a color code reflecting the strength of the underlying evidence. The same report uses rubrics to show how the evaluation team scored the performance of the program in different areas, and annotated almost every argument made with references to specific interviews and other sources of information. The latter, however, risks creating a type of analysis where the emphasis is mainly on reporting what informants and documents have stated rather than adding an additional layer of analysis in which different information sources are considered in an integrated way. The GIE report is less systematic in incorporating rival or commingled explanations in the causal analysis and tends to assume that the observed changes can be linked to the program (although there are some warnings that this is not necessarily the case). The OC evaluation report refers to some rival explanations but not systematically. The DW4W and SD4ALL evaluations, on the other hand, use contribution analysis for a limited number of causal links and add a reflection on rival explanations for the observed changes, resulting in a more comprehensive contribution story.

All in all, the different case studies provide a rich, critical and convincing story of how Hivos and its partner organizations made progress across different domains of change of the CAC program. While there are some differences in the depth of the causal analysis and the engagement with structural drivers underpinning some of the observed changes, all the members of the ERG were positive about the quality of the analysis by each of the four evaluation teams. Along similar lines, the Hivos evaluation managers and the program managers were largely positive about the added value and quality of the case study work.

Do the evaluations achieve sufficient breadth to cover an assessment of the program as a whole?

The thematic evaluation reports all have sufficiently strong argumentation for most of the claims made, but the extent to which this is done in a systematic and transparent way differs. For example, the GIE evaluation report assesses the eight country programs against the different sub-evaluation questions and provides short examples for all the countries in the evaluation. Aside from a detailed analysis of the reported outcomes, the DW4W evaluation collected basic information about the experiences with capacity development for all partner organizations by sending out a survey.[11] The SD4ALL evaluation covered four of the five partner countries through the case studies as well as the international activities of SD4ALL and therefore had a comprehensive scope in its assessment. Finally, the Open Contracting evaluation achieved breadth in its analysis by assessing the thematic case studies across the whole portfolio of the program.

11. The response rates was acceptable (around 60%), and the findings gave insights in overall patterns in how capacity development was perceived by the partner organizations.

© Judith Quax/ENERGIA

Again, while there were differences in how the program-wide assessment was done, each evaluation team devoted substantial time to grasp and assess the overall program. Programs with a smaller number of partner countries (SD4ALL) or a smaller geographical reach and more consistency in the type of partners and settings (DW4W), were easier to assess at the overall program level. Across the four evaluations, the outcome statements turned out to be key sources of information for the evaluation teams. Participation in international partner events where the evaluators could meet all the partners, definitely helped to get a sense of the dynamics at program level and compare the case study settings with other settings. Finally, the way the evaluation was structured with four different phases, further supported learning at the program level and across the four programs. Substantial resources were invested in the learning trajectory after the completion of the four evaluations. This was a worthwhile investment as it enabled in-depth lessons to be drawn across the four programs and to be discussed during a two-day online event with all the program teams of the Citizen Agency Consortium.

In a follow-up focus group session with three of the four program managers, they all concluded that the synthesis evaluation reports gave a comprehensive and balanced representation of the overall program, resulting in findings and recommendations that could be used to improve the program. The ERG also confirmed the quality of the evaluation reports for this component of the evaluation.

3.3. EVALUATION FOR WHOM?

Evaluations of development cooperation programs have to balance the interests and needs of different actors connected in complex relationships with skewed power dynamics. Scholars have argued that many evaluations can be seen as exercises in upward accountability rather than being useful for partner organizations and beneficiaries at various levels. The Hivos evaluation managers deliberately introduced certain principles in the evaluation ToR to align the evaluations with specific learning needs of the four programs (e.g. joint definition of learning questions, final learning event) and to encourage participatory user-oriented activities in the different phases of the evaluation[12].

In their inception reports, the four evaluation teams were explicit about their aims to engage with all stakeholders in the program, and to balance learning and accountability agendas. The learning questions were actively integrated in the case study selection and the evaluation frameworks. The final reports covered four of the five standard OECD-DAC evaluation criteria, and also referred back to the findings on the learning questions. Some discussion emerged between the OC evaluation team and Hivos on the need to add additional learning-oriented activities to the evaluation. The OC evaluation team wanted to trigger double and triple loop learning processes among the partners by inviting them to keep learning journals throughout the evaluation. The Hivos team was not in favor of this to avoid a situation where different program stakeholders would be learning at different speeds, but also to avoid creating an additional burden on an already heavy evaluation process.

The main factors facilitating engagement and learning within Hivos and among the project partners were the following:

Working with an external reference group[13]– The evaluation managers decided to establish the ERG early in the process in order to proactively build in possibilities to adjust monitoring approaches and tools for data collection and analysis, where needed, in anticipation of the end-term evaluation's criteria and requirements. The selection and recruitment of the three ERG members took place in early 2017, combining the criteria of evaluation expertise with subject matter expertise (Lobby and Advocacy and Capacity Development/Assessment). Their main roles were to provide methodological support by bringing in critical perspectives and feedback on proposed steps or approaches. This means that their input and feedback was on invitation of the evaluation managers. Beyond this, however, the ERG members' ample and broad experiences with and perspectives on development cooperation and its history meant that discussion and exchanges often extended beyond the original topic of a meeting. This has benefitted the M&E process in many substantial ways.

Some examples are:

- A shift towards more narrative and dialogical approaches to capacity assessment (CA): although this has not been followed by a consistent shift in all programs and regions, the shift in emphasis did signal the importance of dialogue in CA in support of partnership relations;

- The appreciation of more flexible and adaptive ways of working with theory of change, among others: this supported the use of ToC for the purpose of adaptive management in two programs;

- The appreciation of learning in its many levels: this strengthened the attention to mutual learning in particular, for example in relation to adaptive management and partner ecosystems.

Participation of evaluators in international events with partners – In both the SD4ALL and DW4W evaluations, evaluators had the opportunity to participate in an international event with the partners. While the evaluators' participation in these events was initially not programmed, the evaluators and the corresponding program manager saw this added value as critical. There was a general perception that this: created trust between the partner organizations and the evaluation teams; created ample time for informal moments for exchanges on and observation of more sensitive/complex issues, and contributed to strengthening ownership of the evaluation process beyond Hivos. All informants felt that this should be a standard activity at the start of an evaluation process.

Learning questions and learning event in September 2020 – The program teams, ERG members and evaluation managers were positive about the efforts to identify learning questions with the program teams in anticipation of the evaluation. This brought focus and ownership to the process. As an afterthought, several program managers did feel that they should have been stricter in limiting the number of learning questions to make the evaluation process lighter and more in-depth.

Expertise of the evaluation teams – All four evaluations teams consisted of experienced professionals with three out of four teams having a long track record in multi-country evaluations. Also, the topical expertise was perceived to be strong. This again contributed to building trust and engagement throughout the evaluation process.

Robust and well-resourced evaluation process – The program managers and ERG members acknowledged the fact that this evaluation process was well-prepared, well-guided, adaptive and flexible and that the necessary resources had been mobilized to build a robust and credible evaluation.

At the same time, the following factors were found to be obstructing engagement and learning:

COVID-19 – The evaluations were affected in differing degrees by the COVID-19 crisis (see section 4.5)

End of the program – The end of the Dialogue and Dissent funding channel after only one round of funding, implied that Hivos had to make drastic changes to its programming and partner portfolio, and staffing. None of the four CAC programs were going to be continued. Several program staff members were losing their jobs, a majority of the partner organizations were not going to be retained in new programs. At the time of the final evaluation, this information was gradually filtering through to the different levels, affecting people's morale and motivation to engage and learn to an extent. While none of the evaluators complained about a lack of motivation, it is clear that many of the lessons learned could not be used directly in follow-up programs, aside from the generic, more process-oriented lessons. In this context, some program managers had second thoughts about the mid-term review which had been designed as a light and mainly internal reflection and learning process. In their view, a full-fledged evaluation with external assessors would have brought out some of the learning points and recommendations at an earlier stage when they could still be incorporated in the review of the running program.

Heavy evaluation process – The downside of the comprehensive evaluation process was the fact that it demanded a lot of time and resources, and there was a genuine concern that many of the findings and lessons might not be picked up by the relevant actors. Although there were interactions during the inception phase of the programs, a partnership survey and a 'light' mid-term reflection, the evaluation was the only major CAC-wide exercise and therefore confronted with the complexity of the four programs and the diversity of themes, approaches, instruments, contexts, actors, dynamics, and interests. An overall lack of coherence across the programs made it virtually impossible to do justice to all the learning questions and interests of all the actors without the comprehensive exercise. This also demanded more time and resources than anticipated. In some regions, the evaluation exercise was experienced as rather centralist, steered by Hivos in the Netherlands. However, this was possibly more a consequence of the divergence of the four programs, brought together under a rather loose umbrella concept, rather than a weakness of the evaluation methodology itself. Several program managers argued that the evaluation could have focused on a smaller number of learning questions and fewer OECD-DAC criteria.

Hivos organizational practices – When reflecting about obstacles to learning, recurrent points of critique towards Hivos revolved around two main issues. First, there was an overall perception of a gap between the discourse on more equal partnerships between Hivos and its partners on the one hand, and the management practices on the other. In the CAC program, Hivos moved away from its earlier predominant role as a funding agency to being an actor that engages in joint L&A activities with its partners. Also, capacity development activities were adjusted so that they would be less top-down and more demand-driven. These efforts were acknowledged in the evaluation reports and by program managers, but at the same time, many of the gains made were countered by the increasingly strict and heavy financial due diligence procedures. The four programs struggled with this, especially the smaller partner organizations who could not comply with all the requirements. There was also a real problem of stop-and-go dynamics, as the partner funding periods were too short and continuity of the funding was conditional to extensive financial and narrative reporting. A second problem related to silo dynamics inside Hivos which block learning across programs and actors. This hampered learning across the four programs, but even more so between the CAC program and other Hivos programs. At the level of SP

12. At the timing of writing this reflection note, there was no opportunity to engage with Hivos' multiple partner organizations about their experiences with the evaluation. It was, however one of the main topics of a focus group discussion with three of the four program managers in December 2020. It also emerged explicitly or implicitly as a topic in the evaluation reports.
13. A requirement of the MFA's Grant decision.

CAC, the overarching bodies, such as the project team and the steering committee, hardly played a role above the operational level. Over the course of time, project team meetings revolved more and more around pressing financial and organizational matters, leaving little space for joint strategic and learning-oriented moments across the four programs.

3.4. EVALUATING IN TIMES OF CORONA

The four evaluations were rolled-out in the middle of an unfolding corona crisis. How did this affect data collection and sense-making? What kind of dynamics did it trigger, block or reinforce? What might be the long-term effects of these experiences on evaluation policies and practices? The CAC evaluation was affected by the corona pandemic in several ways. Two evaluations managed to (almost) conclude the fieldwork before COVID-19 made travelling impossible. In one evaluation, only one of the missions was completed, and in the fourth evaluation, all the case study visits had to be cancelled. This affected data collection directly as it is more difficult to do in-depth interviews online, especially when it is further complicated by poor internet connections. It also affected the time available for evaluation indirectly as many program staff and evaluators had to take-on extra tasks due to having children at home (schools closed).

The biggest victim of COVID-19 was the learning, as learning-oriented activities during the fieldwork, workshops and other collective sense-making events were cancelled due to time pressure and technical limitations.[14] This is likely to have affected the 'process use' of the evaluation during the fieldwork as collective sense-making and learning spaces were lost. In addition, for the online case studies, in the absence of most of the 'in-between' moments (when driving to a project location, waiting for an interview, etc.) the evaluators had limited opportunities to test emerging findings, ask further clarifications, etc. Due to COVID-19, there was also substantially less time for interaction with final beneficiaries, which means that the evaluators have to navigate with a more superficial understanding of the context.

However, considering these important hiccups, it is remarkable to observe that none of the evaluation teams concluded that the data collection was heavily compromised. It seems, however frustrating online interviews and focus groups with poor internet connections can be, that the evaluators still had meaningful exchanges with a variety of stakeholders. In countries without a local lockdown, working with local evaluators that could do some basic data collection also helped to improve interactions and deepen the understanding of the contexts. As such, the external reference group did observe differences in the evaluation reports that included fieldwork (before the pandemic) and the other case studies, but not in ways that it questioned the relevance of a strong online component to evaluations. As a consequence, it is likely that, even after corona, evaluations will increasingly have online interactions as an accepted component of evaluation work. COVID-19 is also likely to accelerate an over-due process of increasing the use of local consultants in all the steps of the evaluation process.

14. An international synthesis and learning event with the four programs did come through in September 2020, with good participation and important insights for future Hivos programming.

© Judith Quax/ENERGIA

04 DISCUSSION

Evaluations that seek to balance different purposes may end up serving neither of them sufficiently well.
The argument is often made that there is no trade-off between learning and accountability purposes. When it comes to the practice of allocating scarce evaluation resources, however, this is more complicated. To be acceptable for donor/upward accountability, an evaluation must cover a broad section of the program under review, data collection and assessment is expected to be done by independent evaluation experts, and preferably, all the five OECD-DAC evaluation criteria should be covered. These three conditions put significant claims on the evaluation resources as well as on the evaluation process. This again limits the possibilities for delving deeply into specific learning topics through case studies or other in-depth inquiries. While complying largely with the ministry's evaluation guidelines, this evaluation has challenged some of these assumptions. It has pushed an active learning agenda by centering the four evaluations around a set of context-specific learning questions. It has also encouraged the evaluation teams to work with available monitoring data as much as possible, and to invest a substantial amount of resources into case studies and learning-oriented activities. A major insight emerging from this exercise is that, if donors would attach more 'accountability value' to the reporting by grantees on the basis of their own monitoring data, then this would free up evaluation resources for the purpose of in-depth learning.

'Results' of advocacy work cannot be fully appreciated by M&E approaches that focus mainly on capturing outcomes.
It has become increasingly clear that mainstream M&E approaches do not do justice to the complexities of advocacy practices. One key stumbling block of applying M&E tools that put all the energy into capturing outcomes is that advocacy is not a linear process, and change is rarely caused by one actor only. Advocacy, as it has emerged in a study on the use of Theory of Change, is a deliberately interactive process, in which advocates intentionally, but not always publicly rope in others to open doors, twine and mesh knowledge and personal relations, in proactive or reactive anticipation or reaction to emerging threats or opportunities, among others. Understanding and appreciating advocacy results cannot happen without also understanding how they have come about. By using Outcome Harvesting as the basis for the monitoring of outcomes rather than logframe-based indicators, Hivos and its partners acknowledged this reality. OH has helped to critically reflect on advocacy outcomes, beyond a tick-the-box exercise with indicators, and gain a better understanding of how the programs work. At the same time, the reflection note demonstrates that OH on its own does not provide all the data needed to obtain an in-depth understanding of an advocacy program. To fully understand the relevance of a result (and the efforts to achieve it), M&E approaches such as OH need to be complemented with methodologies that explicitly look to open the 'black box' of advocacy efforts and dynamics that have led to the result.

Contribution analysis in advocacy programs requires thick descriptions and attention to the whole, as well as the parts.
The observation in the previous paragraph also has implications for how contribution analysis should be done in advocacy programs. Outcome statements emerging from the monitoring process provide a basic reflection on the program's contribution to a given outcome, which was essential in orienting the evaluation teams. These contribution claims, however, were too thin in description to serve as the basis for a full causal analysis. They especially lack rich contextual descriptions of the story of change as experienced by the advocates, external accounts of the change process, as well as evidence from secondary data. In addition, while the OH methodology does encourage considering other actors/factors in describing the contributions to an outcome, there is always a risk that a program team has some level of tunnel vision. Considering the complexity and multi-actor nature of social change in an advocacy process, evaluators have the responsibility of performing an integrative analysis, which considers both the whole system of actors and factors, as well as the relative role played by them. The SP CAC evaluation demonstrates the necessity and relevance of case study research and/or qualitative data collection instruments, such as Narrative Assessment, to develop thick and integrative descriptions to validate the outcome statements and build a credible contribution story (or refute it, if necessary).

Maintaining an enabling policy framework for learning-oriented evaluation is key.
At the beginning of this note, we made reference to the deliberate choice of the Dialogue and Dissent framework to move away from managerialism and towards a transformative approach. This allowed consortia to adopt adaptive management practices based on theories of change. Although it is formally a successor to Dialogue and Dissent, the current 'Strengthening Civil Society' framework appears to contain aspects that signal a partial return to the managerialism of old, such as the insistence on target setting and – for some thematic areas – the request to align with stricter results frameworks.

ANNEXES

All reports produced for the SP CAC end evaluation can be found here:
https://www.hivos.org/end-term-evaluation-of-the-citizen-agency-consortium/

Reflection on the Evaluation of the Citizen Agency Consortium Advocacy Programs

KU LEUVEN
HIVA
RESEARCH INSTITUTE FOR WORK AND SOCIETY

Hivos
people unlimited