The effectiveness of support to lobby and advocacy: synthesis of evidence from Dutch MFA programme support

Hugh Sharma Waddington, Hikari Umezawa and Howard White

15/04/2023

Contents

Executive summary

There has been a shift in official development assistance towards programmes that aim to tackle the root causes of poverty through improving the institutions of development and circumstances in which policy decisions are taken, including supporting lobby and advocacy (L&A) by civil society. From 2016 to 2020, the Dutch Ministry of Foreign Affairs (MFA) provided over Euros 1 billion to official development assistance programmes for L&A in low- and middle-income countries and at the regional and global levels. This support was provided through the Dialogue and Dissent (D&D) Programme and the Sexual and Reproductive Health (SRHR) and Rights Partnership Fund.

There are challenges in evaluating L&A, relating to the size and complexity of the programmes and the fact that programme objectives are often to change policies or practices in a single institution like a Presidency or Parliament. Many L&A programmes are therefore not amenable to traditional impact evaluation methods drawing on 'large n' approaches. But 'small n' approaches are available to measure the effectiveness of L&A which use the programme theory of change at their core. Thirty-two of the MFA programmes were independently evaluated using 'small n' methods. We conducted a meta-evaluation of this evidence, to assess the programmes' achievements and the credibility of the evidence on effectiveness, in order to provide guidance about methods for evaluating L&A.

All programmes aimed to engage in community-level and policy-level debates with government or private sector actors, and thus improve policies and policy implementation. They did so by supporting the capacity of civil society to engage with L&A and helping to build partnerships. Programmes with SRHR objectives also aimed to change attitudes and improve service delivery and use. Well over 1,000 outcomes were reported in the evaluations, around half of which were on policy engagement, and over 80 percent of the changes in outcomes were found to be positive. However, assessments of the contribution of the programmes to the changes were not provided in many of the evaluations, which made it impossible to determine if the changes that occurred had been due to particular interventions implemented, and, if so, what was the strength of the contribution. There appeared to be a clear bias in the evaluations towards reporting outcomes that were achieved.

In our assessment, the minimum criteria for confidence in an evaluation, of whatever type, is one that (1) clearly defines its study design, (2) adequately conducts it, and (3) reports how the study was designed and conducted appropriately. However, we believe that evaluations of effectiveness questions should go further than this by (4) attempting to address possible sources of bias in outcomes being collected with reference to an explicit theory of change, (5) addressing sources of bias in causal claims being made through a sampling strategy that includes informed independent information sources and other known sources of bias in reported outcomes, and (6) incorporating approaches that can measure and validate contribution relative to other programmes and relevant contextual factors.

A number of studies had a clear evaluation design at the outset, but provided little description on how the planned design was actually implemented; in such cases we were unable to assess the credibility of the evidence.

Where the evaluations had an explicit method to measure effectiveness, the majority used Outcome Harvesting (OH), which is a method suitable for obtaining information about possible intervention effects from participants or programmes staff. In some instances, other approaches like Contribution Analysis (CA) were used or a combination of OH and CA. The approaches tended to be conducted appropriately, according to best practice manuals, but we identified important areas where evaluation design, conduct and reporting could be improved. These included: providing more information about interventions occurring at the country and grassroots levels, to avoid the problem of "missing beginnings" in the theory of change; justifying the choice of sample and avoiding "omitted informant bias" from those who may not have participated in the programme but who might have had different, but informed, perspectives about its achievements; and addressing predictable sources of bias in establishing effectiveness relating to alternative factors contributing to change, and respondent and evaluator biases. The reporting of outcomes could also have been clearer in many cases, by explicitly using the theory of change to link programme inputs and activities with outputs and outcomes, especially at the country and grassroots levels, and to ensure that the full range of possible outcomes were evaluated, not just those that resulted from effective strategies.

Outcome Harvesting is a practical way to engage programmes staff and participants, and was commonly incorporated in MEL systems in the L&A programmes evaluated. For immediate outcomes relating to capacity building, it would be useful to incorporate as a validation method, more objectives assessments of knowledge and practices of CSO staff before and after the programme was implemented. For credible assessments of effectiveness of programmes in achieving outcomes further along the causal pathway, evaluators and evaluation commissioners should ensure that both successes and failures (outcomes), and alternative explanations, or contributions, by external actors to the achievements (programmes), are measured.

We also believe that evaluations of L&A programmes – and 'small n' impact evaluation more generally – can be improved through clearer guidance including on the design of evaluations, ensuring they engage with outcomes that are achieved and those that aren't; guidance to improve conduct, especially to address common sources of bias; templates indicating what aspects should be included in evaluation reports; and checklists (to be submitted with reports) where evaluators can indicate compliance with best practices.

Abbreviations and Acronyms

CA	Contribution Analysis
CJP	*Justice et Paix*
CSO	civil society organization
D&D	Dialogue and Dissent
EQ	evaluation question
FGD	focus group discussion
FGM/C	female genital mutilation/cutting
GAGGA	Global Alliance for Gender and Green Action
GoP	Government of the Philippines
HIV	Human Immunodeficiency Virus
IOB	Policy and Operations Evaluation Department
L&A	Lobby and Advocacy
KII	key informant interview
MAPP	Method for Assessment of Programmes and Projects
MEL	monitoring, evaluation and learning
MSC	Most Significant Change
MFA	Ministry of Foreign Affairs
MLT	middle-level theory
OH	Outcome Harvesting
OM	Outcome Mapping
PITCH	Partnership to Inspire, Transform and Connect the HIV Response
PT	Process Tracing
QuIP	Qualitative Impact Protocol
SOMO	Centre for Research on Multinational Corporations
SRHR	Sexual and Reproductive Health and Rights
STI	sexually transmitted illness
ToC	Theory of change

Acknowledgements

We thank the many NGOs and evaluators that responded to our request for feedback on the coding. We are grateful to our colleagues in the External Review Group – Karen Biesbrouck, Karel Chambille and Jos Vaessen – for advice and comments on the draft inception and final reports. We also thank colleagues from the Dutch Government who commissioned and supported this study – Ferko Bodnar, Caspar Lobbrecht, Rob van Poelje, Cobi Mars and Noortje Verhart. The views contained in this report, and any errors, are solely our own.

Hugh Sharma Waddington, London School of Hygiene and Tropical Medicine and London International Development Centre

Hikari Umezawa, Campbell Collaboration

Howard White, Global Development Network, Campbell Collaboration and Lanzhou University

Chapter 1 Introduction: evaluating advocacy in international development

1.1 Background

This study is about the so-called shift from the funding of anti-poverty programmes that address the specific constraints faced by individuals and communities (e.g., nutrition, credit, water and sanitation), to those that tackle the root causes of poverty at a societal level, relating to imbalances in decision making power and institutionalised inequality. There is great interest in evaluating the effectiveness of programmes that aim to address societal level problems, but they frequently pose challenges to evaluators. The effectiveness of some societal programmes may be evaluated using traditional experimental and quasi-experimental methods, where there are sufficient numbers of participants for 'large n' statistical designs. In many other cases, however, such as where programmes are aiming to influence decision making in a particular institution (like a parliament or a presidency), evaluating the effectiveness of development programmes and projects requires appropriate methods. These use theory-based approaches to articulate the causal pathways thought to operate to produce outputs and outcomes from the programme inputs and activities, and ideally incorporate some method of testing for alternative causal claims, such as those operating because of other programmes happening or existing capacities.

The Netherlands Ministry of Foreign Affairs (MFA) support for Lobby and Advocacy (L&A) is provided through the Dialogue and Dissent (D&D) and Sexual and Reproductive Health and Rights (SRHR) Partnership Fund policy frameworks. These programmes provided Euro 1.16 billion to consortia led by international non-governmental organisations (INGOs) to support L&A in low- and middle-income countries (L&MICs) in 2016-2020. Examples of D&D programmes are "Freedom from Fear" led by Pax (Euro 50 million), Global Alliance for Green and Gender Action (GAGGA) (FCHM, Euro 32 million), Partnership to Inspire, Transform and Connect the HIV response (PITCH) (Aidsfonds, Euro 50m) and "Towards a Worldwide Influencing Network" (Oxfam/Novib, Euro 78 m). SRHR programmes included "Bridging the Gaps" (Aidsfonds, Euro 50 million) and "More than Brides" (Save the Children, Euro 59 million), among others. These are complex programmes with multiple intervention components that aim to build capacity and support L&A at grassroots, country and wider levels, and in many cases catalyse partnership movements for policy change and, in the case of SRHR, improved sexual and reproductive health outcomes. Multiple actors are involved in these initiatives, often with multiple external players also involved in the various policy areas of interest. Moreover, the programmes usually covered a large geographical area and scope.

Evaluation of the effectiveness of these programmes is complicated. The programmes usually aim ultimately to achieve changes in policies at national and international levels, involve multiple interacting local and global partners, and have multiple objectives and sub-programmes operating in different geographies, which occur at the same time as other programmes or factors which can affect outcomes of interest (Gardner and Brindis, 2017). Therefore, evaluations of these programmes necessarily must use a broader range of techniques than has been commonplace in development impact evaluation, particularly methods that can be applied to 'small n' and 'medium n' cases. The methods available in the evaluation toolkit for evaluating 'small n' cases include Contribution Analysis (CA), Method for the Assessment of Programmes and Projects (MAPP), Most Significant Change (MSC), Outcome Harvesting (OH), General Elimination Methodology and Process Tracing, among others (White and Phillips, 2012; Vaessen et al., 2020).

The programmes necessarily also operated as partnerships, and often aimed to partner with other lobbying bodies, hence it was usually more appropriate to think about contribution, rather than attribution, when assessing effectiveness. "Pushing a car" is a useful analogy when thinking about contribution when multiple partners are supporting and implementing a programme. The idea is that one person can't push a car, two can push it better and three can push it fairly easily. But a fourth person is not needed to push the car, in fact they might get in the way. This exemplifies the principles of necessary and sufficient conditions, which are important components of some types of theory-based evaluations of causal relationships like CA (Mayne, 2020). Thus, with two or three people pushing each is necessary but neither sufficient. The fourth person is neither necessary nor sufficient – they are redundant. Furthermore, the car could also move by being towed, or putting petrol in the engine, or whatever.[1]

An example is the Philippines' Sin Tax, which was a policy to increase excise on alcohol and tobacco. An evaluation of the World Bank's country assistance programme concluded that the Bank was instrumental in getting the Government of the Philippines (GoP) to enact the policy and did not mention anyone else being involved (Kaiser et al., 2015). However, a case study by Harvard (Madore et al., 2015) concluded that the policy was very strongly owned and led by the GoP. This included President Aquino and the Minister of Health, who was a surgeon but keen on preventive approaches, plus the Ministry of Health, who together proposed the Sin Tax to subsidise the health insurance scheme. GoP and the Asia Foundation commissioned a number of studies on the Sin Tax. At one point the case study lists supportive partners, one of whom was the World Bank, although this was one of the only times the Bank was mentioned in the study (Sidel and Faustino, 2019). So, in other words, this was domestically driven policy supported by a wide range of local and international partners. In the example here, the World Bank was the fifth or sixth person pushing the car and so was unlikely to have played a role as significant as stated in the 2015 evaluation.

1.2 Purpose and scope

This study aims to support the Policy and Operations Evaluation Department (IOB) in assessing and summarising evidence on aid effectiveness by synthesising evidence on the effectiveness of Dutch support to L&A via the D&D and SRHR funds. The full list of programmes funded and evaluated under the policy frameworks and the lead NGO partners is presented in Annex 1. We collected the findings from, and assessed the methods used in, evaluations of effectiveness in all 32 external end evaluations, including 25 for Dialogue and Dissent[2] and seven for the SRHR Partnership Fund.

We collected the findings of these evaluations. We developed middle-level theories (MLTs) for D&D and SRHR programmes. We also assessed the strength of the causal claims made in the evaluations and summarised the findings from those evaluations in which we have at least moderate confidence in the causal claims. The evaluations were assessed using a coding tool that was developed and piloted specifically for the purpose of the study. It has been observed that many evaluations do not have an explicit methodology, with the methods section describing data sources only (White and Phillips, 2012; White, 2022). This observation does not discount the importance of data collection as a part of the evaluation design. Box 1 presents terminology which distinguishes between evaluation design and evaluation methods.

The objectives and evaluation questions are shown in Table 1, which aligns the questions with the related objectives. Table 1 also gives a brief statement of the approach to answer each question, which was presented in the study inception report.

It is important to recognise that the evaluations we have assessed were commissioned to address the IOB Evaluation Quality Criteria that existed at the time the evaluations were designed. The assessment we have done, based on IOB's updated evaluation quality criteria, incorporates items reflecting current best practices in the field of evaluation. These criteria are compared in Annex 2. The revised IOB guidelines on which this review was based, which are an extension of the earlier guidelines provided to the evaluators, were not available to the evaluators at the time the studies were undertaken. Therefore, this assessment is not intended in any way as a 'performance review' of those conducting those evaluations. Our intent is to learn from the methods which have been used synthesise the evidence about the effectiveness of L&A.

[1] So, there might be insufficient but necessary parts of a condition that is itself unnecessary but sufficient (INUS) for the occurrence of the effect (Mayne, 2012).

[2] The original remit was to evaluate 25 D&D programmes and 7 SRHR Partnership Fund. However, one programme from D&D programme, the Citizen Agency Consortium, contained four separate evaluations which were assessed individually, bringing the total to 28; these were Sustainable Diets for All, Green and Inclusive Energy, Decent Work for Women, and Open Up Contracting. One programme from SRHR Partnership Fund (the More than Brides Alliance) contained two evaluations (one on Pakistan, the other on India, Malawi, Mali, and Niger), resulting in eight evaluation reports to be reviewed.

Box 1 Evaluation design: data collection and methods

We refer to the evaluation design as all aspects of how the evaluation will be undertaken. That is (1) data collection, which covers both sampling (from whom data are collected) and data collection instruments and approaches (how data are collected), and (2) methods, which are approaches to data analysis.

Since the focus of the work is on effectiveness, we are especially interested in methods used to support causal claims. White and Phillips (2012) noted that evaluations, and even methods papers, are often silent on the basis for causal claims. Similarly, Vaessen et al. (2020) stated that the basis for causal claims in some studies using the outcome harvesting method is "scientifically weak, with no causal model being explicitly used to assess contributions to outcomes" (p.80). However, there may be good reasons for drawing on approaches like outcome harvesting (OH) in programme evaluations, one of which being that OH is commonly incorporated in standard monitoring and evaluation systems, and as an approach can readily obtain information from stakeholders about the perceived or desired (bottom-up) outcome pathways resulting from programme implementation. Where these harvests can be triangulated with other data to verify the causal claims alongside the programme theories of change, the contribution of the programmes can, in theory, be evaluated.

It is also useful to distinguish the proposed evaluation design from its conduct (how the evaluation is implemented) and how the data collection and methods of analysis are reported. We therefore aimed to assess the strength of causal claims made in particular evaluations and for particular methodologies, and to articulate how they may be designed, conducted and reported to foster transparent inferences about effectiveness.

Table 1 Objectives, evaluation questions and approach

Objective	*Evaluation question (EQ)*	*Approach*
Assess how and the extent to which the methods used in individual evaluations allow for a credible assessment of programme outcomes	EQ1: What evaluation methodologies have been used in the evaluation reports to answer the research questions on effectiveness and, when available, impact (namely, effects on longer-term 'final' outcomes)? Were the proposed methods adequately applied in practice (design, conduct and reporting)?	We describe and assess the evaluation design, which covers both data collection and analysis (methods), with a focus on the methods used for causal claims and their conduct and reporting. The data on the evaluation design were collected through a coding form, which is an elaboration of the IOB criteria.
	EQ2: Are the evaluation methodologies as applied in the 32 reports in line with the updated IOB evaluation quality criteria that focus on evaluation methodology?	We elaborate the IOB evaluation quality criteria for effectiveness, specifying a series of sub-questions for each criterion. These questions were piloted at the inception stage. Coding of each study against these questions was carried out by two coders working independently for 20 percent of the studies (4 D&D and 2 SRHR) with consensus reached through discussion.
Formulate lessons and recommendations with regards to evaluating the effectiveness and impact of lobby and advocacy (L&A) programmes	EQ3: What are the appropriate evaluation methods, and their common characteristics, for evaluating effectiveness, in the field of capacity building of, and working with, civil society organizations (CSOs) for L&A and the L&A outcomes they achieve?	Based on our assessment of the evaluation designs we identify the methods for evaluating capacity building of CSOs for L&A which are deemed to yield the most reliable findings. This is supplemented by additional suggestions from wider reading of the literature.
	EQ4: What were the common characteristics for the less suitable methods to rigorously evaluate capacity building of, and working with, CSOs for L&A and the L&A outcomes they achieve?	Based on our assessment of the evaluation designs we identify the methods for evaluating capacity building of CSOs for L&A (factors relating to design and/or conduct), which are less likely to yield reliable findings for reasons of likely bias.

Objective	*Evaluation question (EQ)*	*Approach*
Synthesise the results (at outcome level) achieved by the 32 programmes	EQ5: Based on the evaluation reports and the assessment of the evaluation methodologies, what can be said about the achieved results of the 32 supported partnerships?	We summarize the evaluation findings with respect to capacity building and the effects of supported L&A activities, developing middle-level theories for D&D and SRHR, for evaluations that were assessed as being at medium or high confidence.

1.3 Initial theory of change for lobby and advocacy

A theory of change (ToC) maps the causal pathways from the inputs and activities provided by an intervention or programme to the outputs produced which contribute to the intended outcomes (White, 2009; Funnel and Rogers, 2011). The ToC will also usually incorporate explicit assumptions underlying the steps in the causal chain.

A ToC has two important roles in an evaluation: (1) to frame the evaluation, and so identify relevant evaluations questions; and (2) to test the theory and so conclude why or why not the intervention is working. Where a ToC is presented, evaluations often fail to report the second of these uses. Most evaluations presented a theory of change for the programme, although the strength and validity of the ToC varied (for example, lack of underlying assumptions, theoretical links, etc), as discussed in the results section. The ToCs presented here for D&D and SRHR programmes are presented to help frame the study. In subsequent sections, we fit these project-level theories of change into middle-level theories (MLTs), based on the existing theories of change. The D&D Programme theory of change is shown in Figure 1, which is the Ministry's own ToC for D&D.

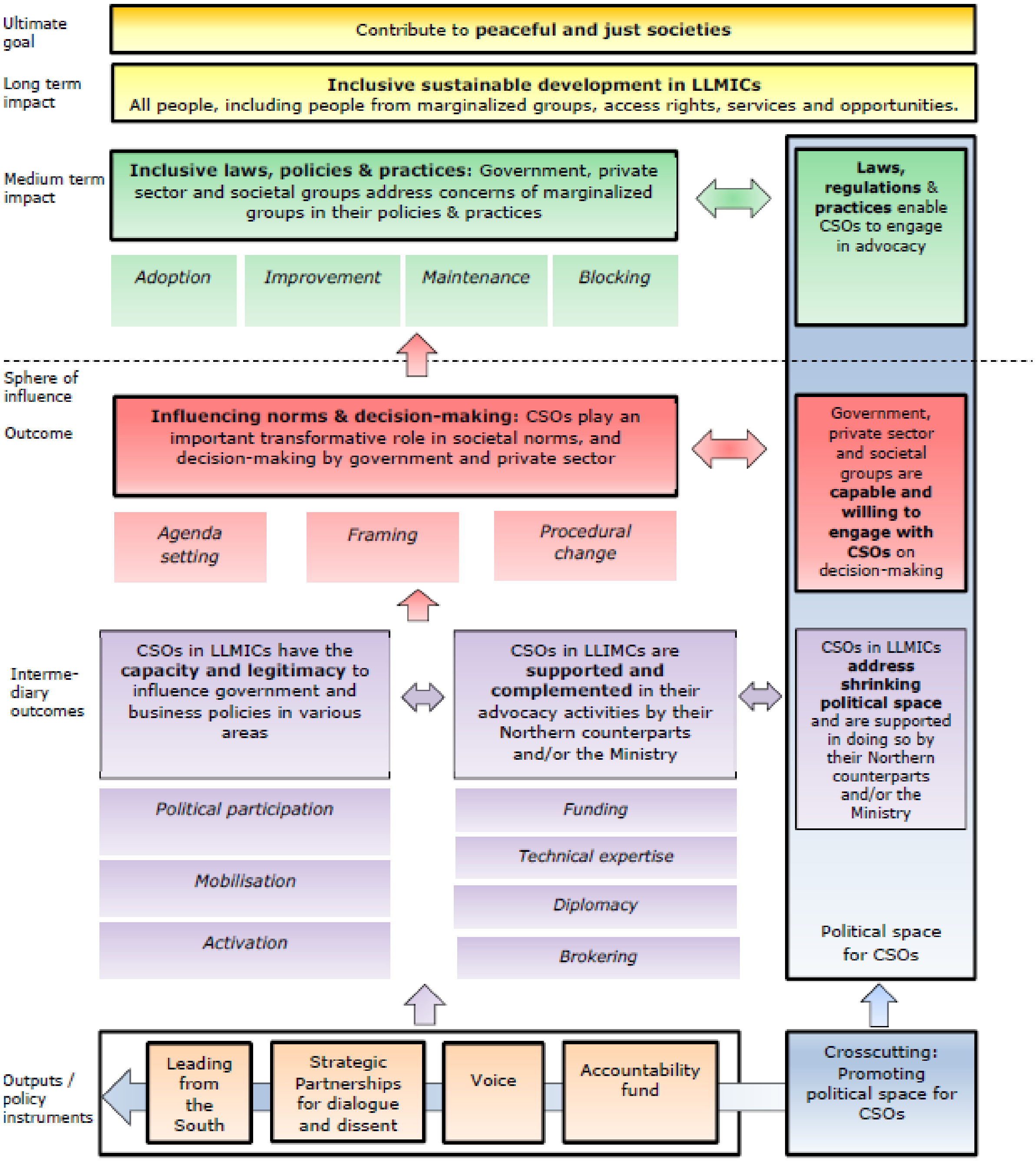
The programme was designed to promote grassroots leadership among Southern partners via strategic partnerships for dialogue and dissent to influence government and other powerful bodies (including those operating in the private sector), to provide voice and funding for projects to improve accountability. These outputs are envisaged to improve the capacity and legitimacy in civil society and civil society organisations (CSOs) to lobby and advocate government and businesses, supported by technically, financially and diplomatically by development partners, including the MFA and other Northern global development bodies. The intention is that these activities, outputs and intermediate outcomes will lead to influence over policy and decision making, with the goal of promoting inclusive laws, policies and practices.

Examples of programme objectives for key outcomes sought included:

- Capacity building: "To build the capacity of southern Civil Society Organisation's (CSO) for 'lobbying and advocacy' (L&A), to enable them to contribute to sustainable and inclusive development, alongside their national and international partners, in order to fight poverty and injustice" (Rainforest Alliance programme evaluation).
- Support to lobby and advocacy: "To strengthen the lobby and advocacy capacities of civil society partner organisations in countries in East & Southern Africa, Southeast Asia, and Latin America as well as at global level, and, together with these civil society partner organizations, on achieving lobby and advocacy goals by influencing policies and practices of market and government actors …. to make more sustainable, diverse, healthy, and nutritious food available to low-income citizens (Citizen Agency Consortium programme evaluation).
- Partnership": "Media and journalists, as independent players in civil society, to constitute a diverse and professional media landscape and function as change catalysts ("No News is Bad News" programme evaluation).
- Policy engagement: "To contribute to a conducive environment in which political and civic actors can effectively influence political processes to advocate for inclusive and equitable social change" (Netherlands Institute for Multi-party Democracy Strategic Partnership Dialogue and Dissent programme evaluation).
- Policy change: "Improved policies, increased investments and better practices on Integrated Risk Management (IRM) at sub-national, national, regional as well as global levels to make… vulnerable people more resilient to crisis in the face of climate change and environmental degradation, enabling sustainable inclusive economic development (Partners for Resilience programme evaluation).
- Policy implementation: "To increase the responsiveness of political parties and parliaments to civic actors in policy processes (Advocacy for Change programme evaluation).

- SRHR outcome”: "To enable people to realize their right to the highest attainable sexual and reproductive health (SRH) (impact), by strengthening health syst”ms" (Health System Advocacy Partnership Programme evaluation).

Figure 1 Dialogue and Dissent Programme theory of change



Ultimate goal
Contribute to peaceful and just societies
Long term impact
Inclusive sustainable development in LLMICs
All people, including people from marginalized groups, access rights, services and opportunities.
Medium term impact
Inclusive laws, policies & practices: Government, private sector and societal groups address concerns of marginalized groups in their policies & practices
Adoption
Improvement
Maintenance
Blocking
Laws, regulations & practices enable CSOs to engage in advocacy
Sphere of influence
Outcome
Influencing norms & decision-making: CSOs play an important transformative role in societal norms, and decision-making by government and private sector
Agenda setting
Framing
Procedural change
Government, private sector and societal groups are capable and willing to engage with CSOs on decision-making
Interme-diary outcomes
CSOs in LLMICs have the capacity and legitimacy to influence government and business policies in various areas
CSOs in LLIMCs are supported and complemented in their advocacy activities by their Northern counterparts and/or the Ministry
CSOs in LLMICs address shrinking political space and are supported in doing so by their Northern counterparts and/or the Ministry
Political participation
Mobilisation
Activation
Funding
Technical expertise
Diplomacy
Brokering
Political space for CSOs
Outputs / policy instruments
Leading from the South
Strategic Partnerships for dialogue and dissent
Voice
Accountability fund
Crosscutting: Promoting political space for CSOs

Source: Kamstra (2017).

The causal relationships in a theory of change are hypothetical. While all relationships in ToCs are probabilistic, the final arrow to policy change or implementation being achieved (distal outcomes) shown in Figure 1 is potentially the most difficult to achieve due to the wide range of variables that affect decision making about policies and their implementation. In other words, ToCs may show necessary conditions, but these are rarely likely to be sufficient, and may be part of an INUS condition; that is, supporting L&A activities by civil society organisations (CSOs) increases the likelihood of policy effects but does not guarantee them. Indeed, Teles and Schmitt (2011) argue that only a small proportion of projects may expect to achieve such success – i.e., supporting L&A is analogous to venture capital in which most investments fail, but when they succeed then the pay-off is high. The implication of this view is that failure to achieve policy change or implementation is not necessarily a failure of the project which should be assessed. One explanation is that it may arise due to the quality of the supported L&A activities. That in turn indicates the importance of evaluating the effectiveness on building capacity of CSOs, or, more generally, immediate outcomes that are further back along the causal pathway which are under greater control by implementers. But failure to achieve distal outcomes like policy changes may also arise due to external factors that can inhibit the project from achieving its intended effects. For example, the project may be well-implemented with activities well-adjusted to the context and implemented with the right expertise (i.e. quality), but other factors (e.g. political instability, weak state capability, external shocks) inhibit the project from having a meaningful effect.

1.4 Approach taken in this study

The approach we developed drew on revised IOB guidelines and best practices in assessment and evidence synthesis. We developed coding tools that aimed to harvest all outcomes contained in the evaluations and to assess the methods used to evaluate effectiveness and the strength of evidence on effectiveness. This methods tool aims to assess dimensions that are considered important in quality assessment frameworks, including the substantiation of findings, application of appropriate methodology, accessibility of reporting, appropriate and inclusive reporting, and analysis of context (Pollard and Forss, 2022). The coding form (Annex 3) was developed collaboratively to reflect IOB's updated evaluation quality criteria (Annex 2) drawing on existing assessment approaches for qualitative evaluation, including CASP (2018) and White et al. (2021).[3] The main points to note are: (1) we included all effectiveness related items from the updated IOB evaluation quality criteria, and some additional items from those guidelines which we thought important for assessing effectiveness; for example, intervention and outcome descriptions; (2) we elaborated the IOB Criteria by breaking them down into several sub-questions, for example by listing possible sources of bias and how they may be addressed (e.g., blinding). Our piloting suggested that breaking down the criteria in this way reduced the need for judgement in applying criteria, thus increasing the reliability and validity of the coding. And, (3) we added questions on reported effects to answer EQ5 about results achieved.

We followed a consultative approach in this evaluation, to promote engagement throughout the study among non-governmental organisations (NGOs) and evaluators of L&A programmes, in order to ensure we fairly reflect their perspective and to have the best chances of engagement with the study findings. Good stakeholder engagement, right from the outset of a project, is considered to be beneficial – e.g., by highlighting flaws in the evaluation questions, design, conduct or reporting – and by creating shared ownership and, therefore, use of the findings. The approach we followed aimed to be consistent with Robert Chambers' (2007) notion of evaluations as the means of "empowering [stakeholders] not [being] extractive". We aimed to create the basis for a good process through an External Reference Group (ERG) comprising users and providers of evaluations assessed in this study. The ERG was brought together in the first month of the project and met four times during the study: at the inception stage to discuss the assessment tool; at analysis stage to discuss preliminary findings from the data collection; at draft report stage to discuss the findings; we also held a public consultation of the preliminary findings at the What Works Global Summit 2022, where ERG members and representatives of the NGOs and evaluation community commented on the preliminary findings. All preliminary findings of the assessments and outcomes were shared with each programme organisation and evaluator, facilitated by IOB, who were able to comment on and challenge the codes that had been assigned. The NGOs and evaluators were allowed 20 working days to provide feedback, sometimes longer. We subsequently updated the preliminary coding for each study, as appropriate, based on the feedback. These revised codes formed the basis of the synthesis findings. We also held meetings with a larger group of stakeholders – the NGOs, evaluation partners and MFA programme managers who supported the programmes we have assessed – to discuss the final report in December 2022.

As part of the assessment tool, we incorporated whether the evaluators provided a positionality statement or discussed their own positionality in relation to the programme or the evaluation participants. In this section, we aim to discuss our own positionality in relation to this study and the evaluations we are reviewing. We have received funding from IOB to assess and summarise the findings from the programme evaluations; IOB were able to

[3] We also reviewed existing literature on the evaluation of L&A to inform the content of the coding forms and our assessment of the evaluation design in the included studies (e.g., Barret et al., 2016; van Wessel, 2018; and Teles and Schmitt, 2011). The literature was identified by Google Scholar and Google searches.

comment on the draft inception and final reports, and facilitated the discussions through the External Review Group and the receipt of feedback on draft coding, but had no role in data collection, analysis or drawing up of implications.

- Hugh Sharma Waddington is a social scientist from the UK, with degrees in economics and environmental health impact evaluation. He has 20 years of post-graduate work experience, and has lived and worked for long periods in India, Rwanda and the USA. As a previous Head of 3ie's London Office and Evidence Synthesis Programme, he has commissioned, designed, led and supported over 100 mixed-methods systematic reviews and impact evaluations of development programmes, which incorporated quantitative and qualitative evaluation methodologies. Much of his methodological research has focused on the design and evaluation quality assessment of evaluations, and he has previously led efforts to develop tools to assess quantitative impact evaluations. He has also designed and conducted primary fieldwork using qualitative approaches in Bangladesh and India, and has a certificate of participation in the residential course on Participatory Methods and Approaches from Praxis Participatory Research, Kerala, India (https://www.praxisindia.org/).
- Hikari Umezawa, who is a research assistant at the Campbell Collaboration based in the United Kingdom, has over a year's experience in synthesis research, data collection and analysis on evidence and gap maps and systematic evidence synthesis, including projects for CGIAR, the Green Climate Fund and the Youth Endowment Fund. She was born and lived in Japan until she moved to France to obtain her bachelor's degree in Economics and Management. After a few year's work experience in the Japanese private sector, she studied MSc Development Economics at the University of East Anglia. As part of her postgraduate course, she was trained on impact evaluations of development programme including mixed methods approaches and was introduced to data collection in qualitative research.
- Howard White is a generalist social scientist from the United Kingdom who has lived in Egypt, Germany, India, Lesotho, the Netherlands and the United States, and spent long periods in Ghana, Sri Lanka, Uganda, Vietnam and Zambia. He has degrees in both Development Studies, and Economics. He has led both impact and process evaluations in a range of sectors over more than 30 years, with field experience in countries across sub-Saharan Africa and Asia. He was the founding Executive Director of the International Initiative for Impact Evaluation (3ie), and is a leading figure in the development of approaches to enhance the relevance and rigour of counterfactual impact evaluations of development programmes. A major focus of his research work in this area has been on the central role of theory of change analysis in development evaluations, whether they use 'large n' statistical methods or 'small n' approaches. He has led mixed methods evaluations employing quantitative and qualitative data collection and analysis for over three decades, and published theory papers on the use of mixed methods in both primary studies and evidence synthesis. He has undertaken work on approaches to evaluating policy influence in the context of bilateral and multilateral agency attempts to influence policy in low- and middle-income countries. He has previously developed assessment tools for qualitative studies, which we drew on in designing the tool presented here.

1.5 Limitations

The main limitation of this study is that we were restricted to a desk review of the available evaluations that were provided to us. No independent assessment was made by us of the evidence claims in the studies reviewed, and the synthesis of causal claims was therefore only as strong as the design, conduct and reporting of the included studies on which the synthesis is based. Thus, a potential limitation relating to development of the MLTs is that we were not direct observers of the change, and project-level stakeholders were not involved in this analysis. Since our synthesis is based on what is reported in the reviewed evaluations, behavioural changes and causal mechanisms covered by the primary evaluator but not reported in the main report of the evaluation were not considered. For example, the total population of outcomes measurable (achieved or not achieved) in D&D and SRHR may be much bigger if they also encompass outcomes from projects and themes outside the samples collected in the included evaluations.

Due to complexity of the coding form and resources available, not all the studies were double coded by the researchers (20% were independently double coded) (Table 1). However, as discussed above, we obtained detailed feedback from the programmes organisations and evaluators on the preliminary coding, which was updated accordingly, and therefore served as an additional quality check on the coding undertaken. In some cases, the organisations provided additional information relating to the evaluation design (e.g., inception reports), which we were able to incorporate in the assessments. However, where these sources were not provided, the assessments were based solely on what was reported in the final evaluations themselves. The assessments were also made at the level of the evaluation, rather than for each outcome that was reported in each evaluation.

1.6 Structure of this report

The rest of the report is organised as follows. Chapter 2 presents an overview of the outcomes data collected and proposes middle-level theories of change for D&D and SRHR. Chapter 3 discusses the approaches the evaluations used to measure effectiveness in the programmes. Chapter 4 discusses implications for the design, conduct and reporting of evaluations of L&A programmes. Chapter 5 presents our assessment in light of the specific evaluation questions we sought to address, and discusses the findings in light of other approaches.

Chapter 2 What did L&A programmes achieve?

Our approach combines the traditional objective-based evaluation with the philosophy of outcome harvesting applied to the identification of outcomes (Wilson-Grau, 2019). The rationale for the use of outcome harvesting may be deemed appropriate as the planned policy outcomes were only identified once the projects were underway; that is, they were so-called emergent outcomes. We developed a typology of interventions for each of (i) capacity building, (ii) support to L&A activities, and (iii) of the L&A activities themselves, as in principle we wanted to assess the relative effectiveness of different approaches. We say 'in principle' as our findings suggested that many evaluations did not provide much detail on the capacity building activities conducted as the country and grassroots levels, and the studies did not attempt to unbundle their assessment of effectiveness of these activities into different components. We used this typology to harvest outcomes from the reports. So, in part, we assessed the causal claims in relation to the objectives identified as a result of outcome harvesting. But there are two caveats here. First, outcome harvesting may neglect objectives which were not achieved. Second, it is arguably possible to define the planned objectives for the capacity development component ex ante. Hence, we assessed capacity building outcomes across all studies.

We used two approaches to synthesise the outcomes. Firstly, as presented in the next section, we drew on the individual outcomes harvested (e.g., capacity building, support to L&A, policy change). Secondly, as presented in section 2.2, we used a used a bottom-up approach drawing on additional detail taken from the programme theories to articulate middle-level theories. The descriptive synthesis is based on 21 evaluations for which we had moderate or high confidence in the findings. We provide clear and transparent procedures for arriving at these classifications of causal claims in Chapter 3.

2.1 Causal pathway analysis

In total, nearly 1,000 outcomes were collected from the evaluations of support to L&A (Table 2). Detailed information about outcomes harvested is presented in this section and also in Annex 4. The outcomes most frequently reported belonged to endpoint outcomes such as community/policy level impacts, while some intermediate outcomes (e.g., skills or capacities of CSOs for D&D programmes) were also commonly reported. In this section we discuss the achievements of the programmes by outcome.

The majority of the outcomes measured in the evaluations were reported to be positive changes: these included measured effects in the areas of capacity development, support to lobby and advocacy efforts, policy engagement, policy change, empowerment and access to SRHR services. Regarding our confidence in the outcomes that we harvested, the evaluations rarely provided a clear explanation of the contribution of the programme activities to the outcome and the strength of the evidence, and where these were reported, they were often rated as 'medium' or 'strong'. Hence, in 69 percent of the outcomes harvested for D&D, and 82 percent for SRHR, the reported contribution was unclear (Table 3).

Table 2 Programme outcomes along the causal pathway

Outcome category	*D&D*		*SRHR*	
	Frequency	*%*	*Frequency*	*%*
Capacity development outputs achieved with partner CSOs	59	6	2	1
Capacity development outputs achieved with other stakeholders	8	1	-	-
Support to L&A activities (or outputs achieved) by partner CSOs	43	4	1	1
L&A activities (or outputs achieved) by other stakeholders	7	1	-	-
Skills and capacities of local partners/ CSOs	133	14	7	4
Spillovers to skills and capacities of other local CSOs	6	1	-	-
Partnerships, coalition building and collaborations with other actors	99	10	3	2
L&A activities by local partners/ CSOs	51	5	8	4
Community-level outcomes	84	9	11	6
Policy engagement	198	21	5	3
Policy change outcomes	110	11	4	2
Policy implementation outcomes	161	17	10	5
SRHR outputs achieved	-	-	6	3
SRHR knowledge	-	-	15	8
Girls' attitudes about SRHR	-	-	13	7
Attitudes of other community members about SRHR	-	-	19	10
Girls' empowerment (e.g., involvement in decision making)	-	-	28	15
Access to SRHR services	-	-	13	7
Access to complementary services	-	-	4	2
SRHR service use	-	-	9	5
Sexual and reproductive health outcomes	-	-	20	11
(Perceived) quality of SRHR service	-	-	4	2
Safety	-	-	2	1
Grand Total	959	100	184	100

Note: - outcome not measured.

Table 3 Summary of outcomes reported in the evaluations

	D&D	*SRHR*
Positive outcomes (%)	89	80
Reported contribution (%)	Unclear – 69 Medium or Strong – 30	Unclear – 82 Medium or Strong – 17
Reported evidence rating (%)	Unclear – 66 Medium or Strong – 33	Unclear – 82 Medium or Strong – 18

Capacity building

There are two types of capacity building outcomes measured in the evaluations: 'immediate outcomes' of capacity development activities (e.g., knowledge), and 'intermediate outcomes' which are the skills or capacities generated (Annex 4). About 70 outcomes were reported on capacity development, 32 of which were from evaluations with high or medium confidence. Most of the medium/high confidence study outcomes (25) were measured as positive effects, and seven were neutral (no increase or decrease). For example, to address lack of appropriate monitoring, documentation and reporting (MDR) skills and understanding of human rights in the Freedom from Fear alliance programme evaluation, the alliance delivered a training programme including human rights theory and MDR tools and skills. At the training, 160 human rights activists and journalists participated, and the training methods were well received. Trainees indicated they appreciated the training content, and their credibility and confidence increased. As a result, they have now adopted the MDR approaches in their work more broadly, which was argued to have led to a more effective L&A activities.

In addition, 140 outcomes were collected on skills and capacities of local partners and CSOs, of which 117 were from evaluations with high or medium confidence. The majority of outcomes measured were positive effects (110), 6 were neutral and there was one negative outcome. For example, the GAGGA programme evaluation explains how the activities contributed to strengthening L&A capacities of the partner CSOs; non-financial support included participatory action research and documentation of environmental threats and their impacts. Evidence suggested that participating in these investigations helped CSOs to acquire new skills and improve their understanding of environmental threats and their impacts on women's right and living conditions. The information obtained from these investigations was used by CSOs to increase the visibility of their efforts and raise awareness on the issues in their communities. Some evaluations drew on the '5Cs framework' for planning, monitoring and evaluation of capacity (Keijzer et al., 2011), including the Citizen Agency Consortium and the Civic Engagement Alliance.

Support to lobby and advocacy

Forty-four outcomes related to support to L&A activities, of which 21 were from high or medium-confidence evaluations. Out of the high/medium-confidence outcomes, 19 measured positive effects. Outcomes measures and data sources used are shown in Annex 4. For example, Health System Advocacy Partnership Programme provided opportunities for CSOs to participate in global and regional forums and assist CSO coordination groups in reviewing and strategizing on relevant policies. The programme also created space for CSOs to influence regional, national and global policies, by helping to develop evidence-based papers and L&A strategies. Another example is from one of the country reports of the "Freedom From Fear" programme evaluation, where financial and technical support (training on transitional justice) allowed a group and network of prison survivors to be established. The group was provided psychosocial support to help them to remain politically active and defend human rights. As a result, this group participated at advocacy events at the EU level in 2019.

Partnerships and partner capacity

Over 100 (102) outcomes were related to partnerships, coalition building or collaborations with other actors, of which 40 were from high/medium-confidence reports. Among these were 39 positive effects, and 1 neutral effects (no increase or decrease in outcome). Outcome measures and data sources are shown in Annex 4. For example, the Health System Advocacy Partnership Programme evaluation stated that the programme successfully brought CSOs, government, the private sector and UN agencies together to improve access to essential medicines and conducting research, which led to evidence-based interventions. This resulted from financial support and technical assistance provided through the programme to strengthen CSO networks and platforms. Spillovers to partner capacity, although considered potentially an important effect of programmes, were not commonly measured (only 6 outcomes were collected).

Policy-level outcomes

The most frequently measured outcomes related to policy engagement and policy change and implementation: 488 outcomes were identified, of which 318 were from high/medium-confidence evaluations, including 299 positive, 12 neutral and 7 negative effects. Outcomes measures and data sources are shown in Annex 4. For example, the Citizen Agency Consortium Open Up Contracting programme evaluation stated that the programme substantially contributed to the decision of the Makueni County Government to adopt open contracting principles and the Open Contracting Data Standard, and to disclose a beneficial ownership registry. The evaluation reported that programme also made a substantial contribution to the County Government taking action on a number of works projects that were delayed. The report provided the following causal explanation: "The Hivos East Africa team lobbied and sensitized the [Makueni County] Governor and his Devolution Ministry with research insights" (p.51), after which the Governor's advisers expressed interest in open contracting by the Open Contracting Partnership. As a result, the Governor established an Open Contracting Technical Team. The Development Gateway and a School of Data Fellow collaborated to develop an open contracting portal, which is now operational and contains up-to-date tender and contract data. Through handing over data to an intermediary, the government became aware of the causes of delays in 6 out of 34 delayed projects.

Knowledge, attitudes and empowerment

These outcomes all relate to SRHR. Eight evaluations (6 are of medium confidence and 2 are low confidence) reported 75 outcomes related to knowledge, attitudes and empowerment, and 57 of them were from medium-confidence studies, 44 of which measured positive effects, 7 were neutral and 6 were negative. For example, the evaluation of "Jeune S3" stated that their activities improved young people's knowledge regarding SRHR across different age groups and genders. The programme activities, in and out of school, radio programme, hotlines, youth clubs, speaking groups and sensitisation campaign delivered information about SRHR. Young people then gained knowledge on, and were able to correct misinformation gained from other sources on, menstruation, contraception methods, HIV prevention and risk of STIs, gender stereotypes, the importance of consent between sex partners, perception of health centre visits and participation in sexually transmitted illness (STI) tests. These changes in knowledge were expected to lead to young people's informed decision about their SRHR.

The "Yes I Do" programme evaluation reported empowerment, measured as adolescents' meaningful engagement to claim their SRHR. In one of the programme countries, the local partner Village Children's Forum (FAD) campaigned against child marriage in their villages, were involved in national advocacy campaigns, and collaborated with other youth groups and with FADs from other villages. They were also represented in the village decision-making meetings. As a result, FAD members and other consulted young people reported they gained the confidence and skills to express themselves and speak in public, and they were feeling more confident in expressing their opinion in the village and their family, thanks to the programme. Knowledge, attitudes and empowerment outcome measures are given in Annex 4.

Community-level outcomes

Nearly 100 (94) community level outcomes were reported, of which 46 were from high/medium-confidence reports. One neutral effect (no progress toward targeted results) and one negative effect were measured, the remaining 44 being positive effects. The list of outcome measures and data sources is in Annex 4. For example, the "Freedom from Fear" evaluation measured the programme's contribution to increasing awareness of the importance of human rights violations among citizens and officials, by providing human rights activists training on human rights theory and effective monitoring, documentation and reporting approaches, which led to a more credible human rights sector and evidence-based advocacy activities. Another example of community level outcomes from the same study included peacebuilding intervention in a country at the risk of conflict. In response to escalation of land dispute and violence between two communities in a territory, PAX visited the area along with a programme partner, who was asked for help by the church. With PAX's financial support and advice, the programme partner implemented a series of activities to facilitate a resolution, including "peacebuilding workshops, public awareness campaigns using radio and community meetings, face to face discussions with influential players and potential spoilers, negotiations with and compensation to the traditional land-owners, and constant engagement with the local and customary authorities." As a result, one of the communities agreed to offer some land to the other, and this was acknowledged by both parties in customary acts. Reconciliation of the two communities has been in progress (a ceremony of reconciliation was held and registration of this change with the authorities envisaged), and the violence in the area had stopped at the time of the programme evaluation.

SRHR service access and use

Outcomes were reported with regards to access to, and use of SRHR services (detailed information on outcomes measured is reported in Annex 4), in 26 cases, with 21 outcomes coming from medium-confidence studies, of which one effect was neutral, 2 were negative, and the remaining 18 were measured as positive effects. The "Bridging the Gaps" programme evaluation provided evidence on positive effects on access to SRHR service: the programme supported a local organisation led by male sex workers through various activities, including production of evidence that show the need for accessible, affordable and friendly service. This evidence strengthened the organisation's advocacy efforts to open a community-led clinic. The organisation also provided a replicable model to supply effective and comprehensive HIV prevention packages for sex workers and other key populations. They involved and empowered paralegals, clinicians/nurses, counsellors as well as community members as peer educators, and they became key players in service delivery. These achievements led to expanded access to HIV prevention and treatment for all key populations. The same report also show evidence on positive effects on SRHR service use among transgender persons: A study showed that violence against transgender people was present in Kenya, and the national government did not show willingness to focus on trans gender issues. In response to this, the Bridging the Gaps programme collaborated with multiple partners contributed to L&A efforts from 2016 to 2019 through different platforms, including hosting government trans conversation on HIV programming. Their activities led to the development of the Transgender Guidelines within key population programming in Kenya. The government health practitioners adopted the WHO blueprint for transgender health care, to make sure gender affirming health care services are readily available. As a result. Government services now recognise Trans men and Trans women, and transgender persons feel safer in government hospitals. The report concludes this brought about an increase in the use of health services among transgender people.

Health outcomes

Twenty outcomes measured related to sexual and reproductive health, with 17 coming from medium-confidence studies, of which eight were neutral, three negative and the remaining measured as positive effects. Outcome measures and data collection approaches are given in Annex 4. For example, the "Bridging the Gaps" programme evaluation showed that in Kyrgyzstan, the number of STIs among sex workers in target area: in the programme area, the only STI service was inconvenient for sex workers because of distance to public transport stop and working hours. In addition. The police cleansing forced them to hide and change their usual place of work, which reduced their access to STI services. In response to this, the programme supported a local partner to purchase a mobile unit to provide STI diagnostics along with some consumables, for free in most cases. The mobile unit also provided HIT testing services and pre/post-test counselling. The report provides quantitative evidence on increase in the number of sex workers who received STI services over the programme period, and decrease of STI incidence among sex workers.

2.2 Middle-level theories

To identify regularities in changes caused by the interventions and draw transferrable lessons that can be applied to other contexts, we developed a middle-level theory for each of the programmes (D&D programme and SRHR Partnership Fund). This analysis was only done for the 21 evaluations at 'medium confidence' or 'high confidence', based on our methodology assessment, presented in Chapter 3.

Cartwright (2020) presents 10 steps that are used to develop MLTs:

1. Specify the overall theory (what the programme is expected to achieve and why).
2. Produce a step-by step diagram of causal pathway.
3. Describe the causal principles at work at each step of the causal pathway.
4. Add support factors to the diagram (identification of "enablers" from evidence).
5. Add derailers to the diagram.
6. Add safeguards against the derailers (if information is available).
7. Allow for causal loops.

8. Specify the expected range of application (intervention context).
9. Draw implications for evaluation questions and for monitoring & evaluation indicators.
10. Draw implications for future programme design.

Steps 8 to 10 were omitted from the analysis because they were beyond the scope of the study (we are evaluating evaluations, and not evaluating programmes).

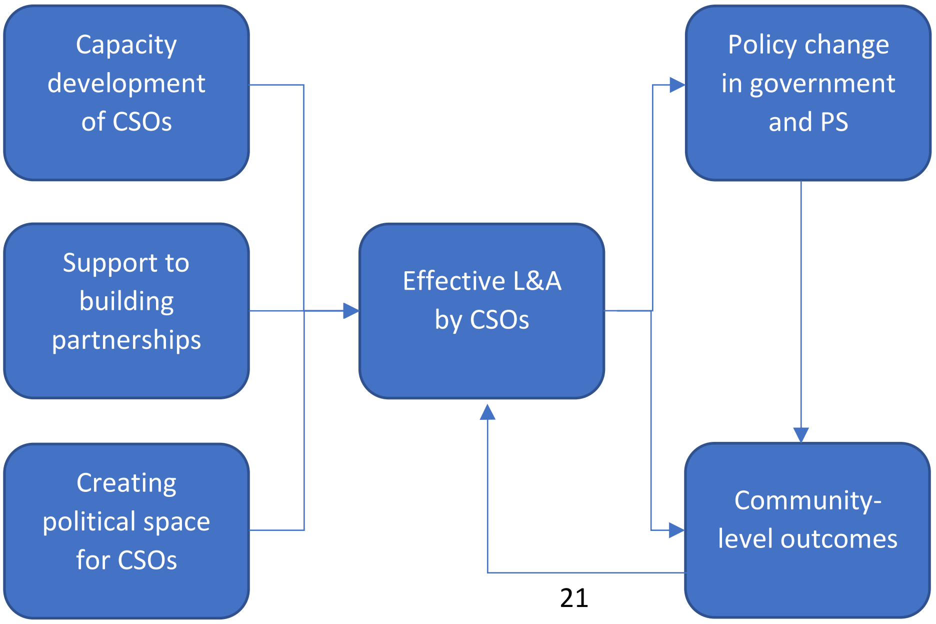
Support to L&A through D&D

Collecting and synthesising outcomes and causal mechanisms reported in 15 evaluation reports (of which two are from the same programme) that we consider as having medium- or high-level confidence, we have developed the following MLTs for the D&D programmes. Support to CSO’s L&A activities lead to inclusive laws, policies and practices for peaceful and just societies, because this supports results in more effective L&A activities by CSOs in the following ways:

(1) Developing CSO's capacity to deliver evidence-based L&A with clear definition of key issues increases their effectiveness.
(2) Developing CSO's strong partnership with other CSO's and key stakeholder increases their effectiveness.
(3) Support to CSO’s L&A by creating political space leads to effective L&A activities.
(4) The increased effectiveness of CSO’s L&A activities lead to the desired policy/community level impacts, because
(5) CSO's activities with enhanced engagement of key actors leads to desired and impactful policy change and implementation.
(6) CSO's L&A with mobilisation of community members and local gatekeepers result in desired community level change, and the mobilised community members get involved in L&A work.

Figure 2 is a visual presentation of middle-level theories for support to lobby and advocacy programmes, provided in more detail in Annex 6. Whilst constructed at a very high level, the figure captures the basic causal process, by which capacity development of local CSOs is expected to enhance their capacity to undertake L&A activities which, in turn, have policy and community-level effects. These are the processes of focus for the evidence synthesis part of this study. The local CSOs are also directly supported in their L&A activities (with regards to partnership building, and creation of political space). There may be other spillover effects which can be (1) internal as supported organisations may undertake other non-project campaigns, and (2) external in other organisations through direct observation, experience-sharing and staff movements (Teles and Schmitt, 2011). The spillover effects were recognised and discussed in some of the evaluations (e.g., under ‘Dissemination of capacity development’). The CSO’s L&A activities also influenced local community, by, for example, affecting social norms and people’s awareness of rights of marginalised people, which then result in community level changes.

Figure 2 Middle-level theory for support to L&A through dialogue and dissent


Capacity development of CSOs
Support to building partnerships
Creating political space for CSOs
Effective L&A by CSOs
Policy change in government and PS
Community-level outcomes

In order for these lessons to be transferrable to another context (Vigneri, 2021), we have identified supporting factors (enablers) and blocking factors (derailers) for each MLT. Enablers are the conditions that were necessary for the desired change to happen, and derailers are the factors that might prevent the change from happening, and hence that require some form of countermeasure. For example, L&A activities with strong engagement of key actors, say private companies in the extractive sector, will lead to better practices in their business operation and encourage them to abide by environment-related regulation or policy. This will be achieved if the CSO successfully engage the most influential companies' key persons (enablers), but their attempt to reach them might cause undesired effect if CSO's approach lacks understanding of the companies' culture – so a safeguard against this derailer will be to organise informal meetings with them to build trust. A fuller list of enablers, derailers is given in Annex 6.

L&A for sexual and reproductive health and rights

The mission of the Netherlands' Ministry of Foreign Affairs in the field of sexual and reproductive health is to promote the universal fulfilment of Sexual and Reproductive Health and Rights. There are four interrelated objectives to achieve this mission: L&A activities for rights, knowledge and attitudes; information and choice of target groups; common attitudes; and improved access to and use of quality reproductive health services. Figure 3 puts these objectives into a middle-level theory, presented in full in Annex 6. The MLT contains the following components:

(1) L&A activities help improve rights and attitudes about SRHR for women, girls and disadvantaged groups.
(2) Rights allow for SRHR choices for women, girls and disadvantaged groups.
(3) Information about rights and services gives women, girls and other disadvantaged groups the knowledge to make informed choices.
(4) Positive attitudes provide the supportive environment for realising SRHR.
(5) Improved access to quality reproductive health services, helps promote their use, leading to improved SRHR outcomes.

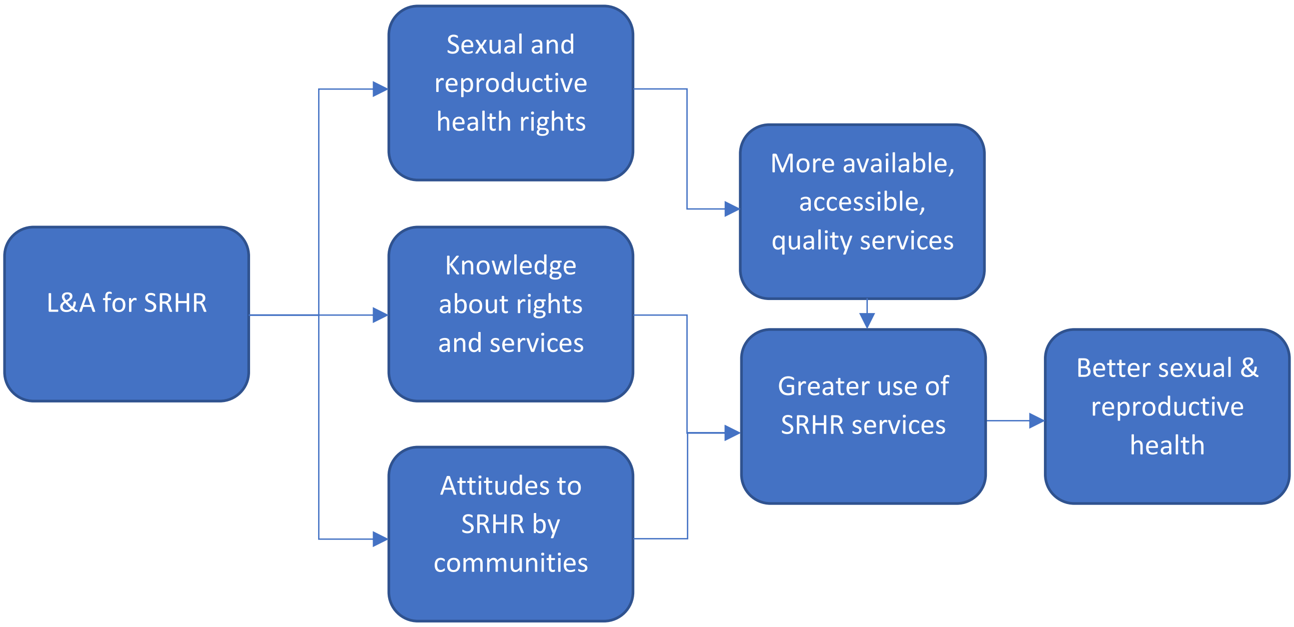
Rights are necessary to allow choice to be possible, for example the right to abortion or the right to refuse early marriage. Rights are of course conditioned by the legal framework which would encompass areas such as legislation against female genital mutilation/cutting (FGM/C). Next, information, together with rights, allows people to make informed choices. However, some of these choices are dependent upon attitudes by community gatekeepers to the SRHR choices available, such as regarding age of marriage, and the ability to practice proscribed behaviours in secrecy or camouflaged by other behaviours, such as FGM/C being practised alongside male circumcision. Choices are also contingent on the availability, accessibility and quality of SRHR services. However, if available, accessible, and of good quality services are present and utilised, consequently better SRHR outcomes are realised.

Enablers and derailers underlying this causal process are:

1. L&A activities are undertaken in partnership with relevant groups, such as women's rights groups in the case of advocacy for sex workers' rights, and use common language among advocacy groups.
2. The communication must be made in a form which will reach, be understandable by and appropriate to, the target group. Communication around SRHR can be sensitive, and so finding appropriate channels through which men and women will meaningfully engage should take into account local norms and values. And health workers need the communication skills to apply this approach.
3. There is a supportive legal framework, including laws criminalising harmful practices, anti-discrimination legislation, prosecutions of those breaching laws such as on child sexual exploitation, and compensation for survivors and victims. Safe spaces may need to be provided where vulnerable groups like sex-workers can obtain information and access to services.
4. There is a supportive policy and practice framework, including guidelines, action plans and treatment protocols practised by health workers and other public sector workers, e.g. community paralegals.
5. The proposed intervention must be attractive to the intended beneficiaries, including those who will use SRHR services and the households and communities in which they live. Promoting demand for contraceptives will be less effective if women – or their relatives – want many children, so the appropriate intervention may need to incorporate messaging around family size, or tackling the factors that make large families attractive, such as high child mortality and son preference.

6. It must be possible and beneficial for the target group to adopt the intervention. Promoting the use of modern contraceptives will not be effective if they are not available, which can be the case especially in rural areas in many developing countries. Or women may be constrained in their use of contraceptives by their partners or other family members.

Figure 3 Middle-level theory for L&A support for SRHR


L&A for SRHR
Sexual and reproductive health rights
Knowledge about rights and services
Attitudes to SRHR by communities
More available, accessible, quality services
Greater use of SRHR services
Better sexual & reproductive health

Chapter 3 What is the strength of evidence on effectiveness?

We elaborated the approach to assessing the strength of evidence used using a consultative process. For each of IOB's updated evaluation quality criteria, we articulated a number of signalling questions on which the criterion was evaluated (Annex 3). For example, we considered the methods used to evaluate effectiveness and whether these were conducted appropriately, as well as the selection of people from whom data were collected, how, and the weight given to different voices. Hence a first question for the methodology was what approach was used to evaluate effectiveness, whether Outcome Harvesting, Contribution Analysis or other? Regarding conduct, a question was whether the evaluation conducted a stakeholder mapping, and, if so, what was the source of the data for that mapping? We subsequently assessed whether the sample of people spoken to was drawn from across the stakeholder map? We attempted to investigate "insider bias" whereby evaluations speak to people inside the project or closely connected to it, but not to those outside the project or those who might even be actively excluded. Evaluation teams may not speak to politicians, religious leaders, trade unionists, traditional leaders, such as chiefs and headmen or women, and journalists, even though these are all important groups of opinion leaders who may be well-informed regarding the issues at hand. Another example is whether the evaluation measured CSO capacity, and, if so, how? If they did so, what method was used to assess whether any changes resulted from project activities, either in aggregate or by component? We therefore also investigated whether assessments of capacity, either that already existing or built and supported through the programme, were made.

3.1 Methods used to evaluate effectiveness

The method, or combination of methods, used in the evaluations of L&A are presented in Tables 4 and 5. The tables show the method the evaluators planned to use, in their methodology section, and method the evaluators said they actually used. Outcome Harvesting was the most commonly used approach, alone or combined with another method such as Contribution Analysis, Most Significant Change (MSC), and Realist Evaluation. Some studies were coded UC (unclear) as they presented how data were collected (e.g., through key informant interview (KII), focus group discussion (FGD), workshop, etc.) but did not describe an evaluation design or method used to ascertain effectiveness (that is, the contribution of the programme to the outcomes achieved). All evaluators used the method they had planned method, with the exception of one SRHR programme evaluation. Contribution Analysis was the second most commonly used approach, either alone or complemented by another method such as outcome harvesting or Method for Assessment of Programmes and Projects (MAPP). In addition, three evaluations of SRHR programmes used a quasi-experimental design (that is, they compared results for participants who received the SRHR intervention with a comparison group which did not).

Table 4 Methods used in D&D programme evaluations

	Outcome harvesting	*Outcome harvesting + contribution analysis*	*Outcome harvesting + MSC*	*Outcome harvesting + realist evaluation*	*Contribution analysis*	*Outcome mapping*	*Outcome mapping + contribution analysis*	*UC*
Planned method	8	6	2	1	3	1	2	5
Actual method	8	6	2	1	3	1	2	5

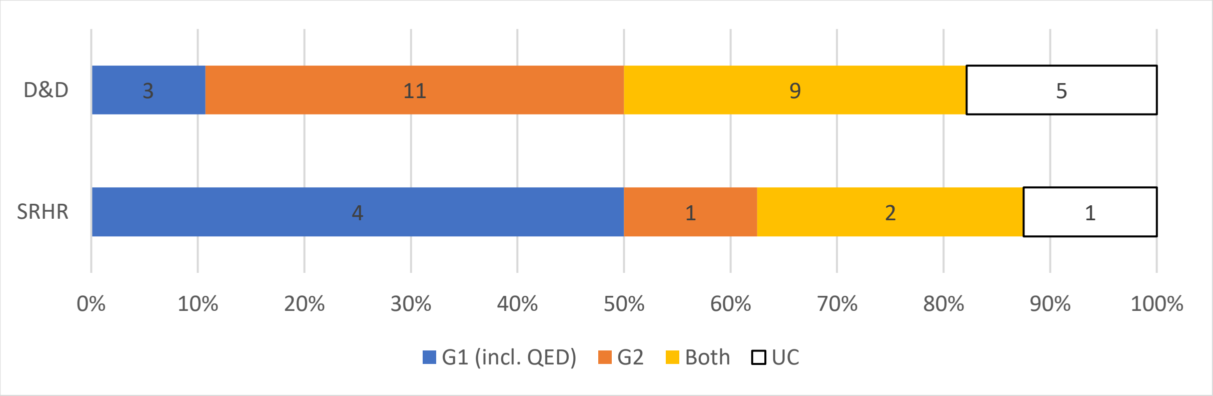
Table 5 Methods used in SRHR programme evaluations

SRHR	Outcome harvesting	Outcome harvesting + contribution analysis	Contribution analysis	Contribution analysis + MAPP	Quasi-experimental	Unclear
Planned method	2	0	1	1	3	1
Implemented method	1	1	1	1	3	1

White and Phillips (2012) distinguished between what they called 'Group I' and 'Group II' approaches to 'small n' impact evaluation. The former explicitly address establishing causal relationships with reference to a theory of change which is tested with reference to a range of evidence sources; the latter are focused on stakeholder views as to what has worked or why, but may incorporate triangulation of data sources to validate the stakeholder views. Stakeholder views and experiences can help shed valuable light on causal processes. Nancy Cartwright has argued that qualitative data can 'vouch for' causal relationships, whereas RCTs can 'clinch' the argument. According to that argument, both Group I and Group II approaches fall into the vouching category,[4] and our data extraction form covered questions which may apply to both, though the degree to which a specific question applies may vary. Both types of approach need to be transparent in the conduct and reporting of their methods and results, which is why conduct and reporting are a major focus of the data collection form.

Figure 4 presents the frequency of Group I (more explicit causal identification, such as contribution analysis, realist evaluation and quasi-experimental design) and Group II (more participatory approach, such as most significant change, outcome mapping, outcome harvesting, MAPP) methods, defined by White & Phillips (2012). It suggests that, in general, evaluations for D&D programme tended to use participatory method or a combination of participatory and causal identification method, while SRHR programme were more likely to be evaluated using causal identification methods.

Figure 4 Classification of evaluation designs by White and Phillips' Groups I and II



We also assessed the conduct and reporting of the methods. We used the following definitions from White and Phillips (2012) and Wilson-Grau and Britt (2013) to assess conduct:

- Outcome Harvesting (OH) involves: 1) gathering data on potential outcomes to which change agent may affect and contributions by change agent; 2) verification through informant review of draft outcomes, usually in workshop, and evaluator assessment of plausibility and coherence; 3) substantiation of outcomes and contributions through additional data interviews; and 4) categorisation and interpretation

[4] In theory, a process tracing approach has the principles in place to enhance rigor in causal inference, and if well-implemented (and with the right data) can 'clinch' the argument.

of outcomes. Similarly, Outcome Mapping (OM) involves: 1) articulating ToC "intentional design" and "boundary partners"; 2) collection of outcome, strategy and performance journals, which may incorporate Most Significant Change (MSC) analysis; and 3) "evaluation planning" (data collection and verification).

- Contribution Analysis (CA) involves: 1) articulating the ToC; 2) evaluating whether intervention activities implemented as set out; 3) chain of expected results (outcomes) shown as having occurred; and 4) other influencing factors ruled out or relative contribution recognised.
- Most Significant Change (MSC) involves: 1) defining domains of change and timeframe; 2) systematic collection of stories from participants about (positive and negative) changes that occurred in their lives in the recent past, enquiries about why the changes occurred and were significant; 3) systematic review of stories of change by stakeholder panels; 4) verification of stories through additional data collection and possible quantification of changes; and 5) comparison of most significant change stories with expected changes in ToC/log-frame.

As noted, OH and CA were usually seen as the most applicable approaches for evaluating L&A. Box 2 presents examples of OH, MSC and CA, where the methods were carried out according to accepted standards, and where, in our view, some aspects of the approach were not adequately reported as having been done. It is worth highlighting here that, even in the first example, where OH was done according to best practice, attempts did not appear to be made to rule out other possible sources of the changes in outcomes that were observed to occur. This is a problem with use of the method of OH alone to evaluate effectiveness, particularly with respect to distal outcomes (e.g., policy change and policy implementation), an issue we return to in the next chapter.

Box 2 Use of Outcome Harvesting, Most Significant Change and Contribution Analysis

One study that applied CA appropriately presented clear information about the evaluation design and conduct. In the first step, the changes taking place were mapped to possible causal claims and stakeholders from whom evidence can be collected. Step 2 involved collecting evidence, including from stakeholders, about the relative contribution of primary (programme-related) factors and secondary (other contributory and contradictory) factors. In the final step, the contribution claims for the programme were defined in light of the contributions of secondary factors. These were subsequently presented in the report using conceptual frameworks which mapped the primary, contributory and contradictory factors for each outcome and country, and indicated which were likely to be the most significant causal claims.

A study that used OH alone followed and reported all of the steps presented above: the data were gathered on potential outcomes and contributions by change agent; these were verified through informant review of draft outcomes, and the evaluators assessment of their plausibility and coherence; the outcomes and contributions were substantiated through data interviews; and outcomes categorised and interpreted. The evaluators used a software visualiser tool, which presented data in graphs, of which copies were included in this report. The evaluators analysed clusters of responses, examined outliers and combined and compared data to answer different evaluation questions. Furthermore, they analysed the harvested outcomes and the results of the exercise during the outcome harvesting workshop where outcomes were mapped in causal pathways and linked them to the ToC and pathways of change, to provide answers to the evaluation questions. But the evaluation did not incorporate a methodological component that aimed to account for possible alternative causal claims.

But another evaluation using OH, which did not describe clearly the methods used or report conducting OH according to accepted methodologies, the evaluators also noted: "The absence of ... the ability to measure capacity-building at the output, ...means that the programme will be challenged to meaningfully evidence what capacities have been developed, which approaches do and do not work, in what contexts, and how sustainable these efforts are" (p.42). In other words, it was not possible to determine whether the programme was building, supporting or using existing capacities of local organisations working in L&A.

In one evaluation that combined OH with MSC, the evaluators compiled the most significant change analyses done by the programmes teams using outcome harvesting. However:

- there was a missing step in the evaluation methodology between the MSC stories harvested and the presentation of findings in the report; in particular, there was no explicit verification of the MSC stories or discussion of how triangulation with interviews/FGDs was used to determine which MSC stories were more or less credible, which the evaluators said was partly due to inability to travel during COVID19;
- similarly, while high level barriers and enablers to effectiveness were discussed, there was limited discussion of alternate explanations for the outcomes achieved at the grassroots level; and, relatedly,

- there were 'missing beginnings' at the grassroots level, so while the programme theory was clearly mapped using the global ToC and data presented on activities and outputs at the high level, there was very limited reporting of activities at the grassroots level to demonstrate temporal precedence – that the actions undertaken by CSOs was related to the actions of the programme itself, hence the issue arose about whether capacity was built or being utilised.

3.2 Assessment of data collection and analysis

The 65 signalling questions that were developed for each evaluation quality criterion (Annex 3) were evaluated used a coding system: 'yes', 'probably yes', 'probably no', 'no' and 'unclear'. We note that 'unclear' meant that there was insufficient reporting of information in order to address the signalling question. This lack of clarity therefore primarily related to lack of information about what was done. While this may have also led to a lack of clarity in our own judgment, the primary evidence on which the assessment is based relates to transparency around conduct and reporting. As is considered good practice in meta-evaluation (Sharma Waddington et al., forthcoming), we developed the assessment tool to avoid incentivising weak reporting on aspects where the conduct itself was weaker. The feedback from the NGOs and evaluators on the preliminary coding revealed that many low scores came about through "insufficiency in reporting" rather than "inappropriateness of approach described". This could be because the final evaluation we reviewed was often a synthesis of country reports.

Subsequently, 35 of the signalling questions were selected to calculate overall and criterion-wise scores and compare the assessments of each study. We used a selection of the questions in the scoring, to avoid double-counting of any factors relating to confidence (the questions included in the scoring are indicated in Annex 3). Scoring of each study for each signalling question can be found in Annex 5. The scores for relevant criteria were added together to reach an overall score for each evaluation. In order for the evaluation to be assessed as having 'medium confidence', the evaluation needed to clearly define the outcomes of interest (Q10.3), report clear effectiveness questions (Q10.6), posit plausible causal mechanisms linking activities to outcomes (Q11.5), adequately report sample characteristics (Q14.6), and use multiple, separate information sources (Q17.1). In order for the study to be assessed as having 'high confidence', the evaluation additionally needed to describe capacity building and/or L&A activities (Q10.2), present a timeline showing cause preceded effect (Q11.4), clearly describe the qualitative methodology used (Q11.7), present a good theory of change (Q13.1), present a stakeholder map (Q14.1), justify the sampling approach used (Q14.5), describe and present the analysis process in sufficient detail (Q15.1, Q15.6), present an evaluation matrix (Q16.1), use appropriate sources of evidence (Q17.4) including those not involved in the programme (Q17.5), triangulate evidence (Q18.1), present alternative possible causal claims (Q19.1), attempt to rule out alternative explanations (Q19.2), attempt to protect against respondent bias (Q19.5) and evaluator bias (Q19.7), and clearly describe how data were collected from informants (Q20.3) and document review (Q20.4). The codes were converted into the scores presented below.[5]

Table 6 presents a summary of the results, with the average score and percentage of possible maximum score for D&D and SRHR programmes, as well as maximum and minimum scores. The table suggests that evaluations for both types of programmes were assessed favourably (scored more than 70%) on criterion #18 (triangulation of results using different information sources); data triangulation was the most common method of triangulation. However, the average score was low (below 30%) for ways to protect against biases (e.g., evaluator bias, respondent bias). Greater challenges were identified for D&D than SRHR, particularly with regard to Criteria #13 (choice of indicator), #14 (sample selection), #15 (appropriate analysis) and #20 (clear description of data collection/analysis). The overall score for D&D programmes was also lower at 39%, compared to 45% of SRHR programme. These findings, and intuition, suggest that it is simply more difficult to design studies to evaluate credibly the effectiveness D&D schemes. Two D&D programme evaluations were assessed as having 'high confidence' overall in the findings. Six evaluations of SRHR programmes and 13 of D&D programmes were assessed as at 'medium confidence'. The remaining studies were assessed as at 'low confidence' in the findings.

[5] The following scoring system was applied: Y=3, PY=2, PN=1, N=0, UC=0.

Table 6 Detailed results of assessments by programme type

IOB criterion	*Criteria description*	***Maximum possible score***	*D&D*						*SRHR*					
			Average score	*%*	*Max score*	*%*	*Min score*	*%*	***Average***	*%*	*Max score*	*%*	*Min score*	*%*
# 10	The research design is clearly elaborated and shows how the research results will contribute to answers to the evaluation questions.	**12**	**8.46**	71%	12	100%	5	42%	**7.5**	63%	12	100%	2	17%
# 11	The methods are appropriate to evaluate effectiveness: attribution and / or contribution	**12**	**5.29**	44%	12	100%	0	0%	**5.63**	47%	9	75%	3	25%
# 13	The indicators or result areas are appropriate to capture the planned results along the different levels in the ToC	**12**	**3.61**	30%	9	75%	0	0%	**5.5**	46%	8	67%	3	25%
# 14	Justified choice of sample, cases and information sources (e.g., choice of countries, projects, organisations and persons)	**9**	**2.71**	30%	9	100%	0	0%	**4.50**	50%	7	78%	2	22%
# 15	The analyses are appropriate, given the chosen research design	**6**	**2.75**	46%	6	100%	0	0%	**4.50**	75%	6	100%	2	33%
# 16	Summary of the methodology in an evaluation matrix and Criteria #17 Sufficient	**3**	**1.36**	45%	3	100%	0	0%	**0.63**	21%	3	100%	0	0%

IOB criterion	*Criteria description*	***Maximum possible score***	*D&D*						*SRHR*					
			Average score	*%*	*Max score*	*%*	*Min score*	*%*	***Average***	*%*	*Max score*	*%*	*Min score*	*%*
	independent information sources													
# 17	Sufficient independent information sources	**12**	**5.68**	47%	12	100%	0	0%	**5.13**	43%	10	83%	3	25%
# 18	Triangulation of results from different information sources	**3**	**2.14**	71%	3	100%	0	0%	**2.13**	71%	3	100%	0	0%
# 19	Discussion of bias	**24**	**4.00**	17%	12	50%	0	0%	**3.38**	14%	10	42%	0	0%
# 20	Systematic, complete and transparent description of the data collection and analysis	**6**	**1.96**	33%	6	100%	0	0%	**2.88**	48%	4	67%	0	0%
# 21	Discussion of the limitations of the evaluation	**6**	**3.29**	55%	6	100%	0	0%	**5.00**	83%	6	100%	2	33%
Total		**105**	**41.25**	39%	73	70%	19	18%	**46.75**	45%	56	53%	37	35%

This section examines the coding in detail for each criterion, drawing on relevant examples. We discuss the assessments for each updated IOB criteria in turn. These include: research design (criterion #10), methods to evaluate effectiveness (#11), theory-based approach (#13), sampling (#14), methods of analysis (#15), evaluation matrix (#16), information sources (#17), triangulation (#18), discussion of bias (#19), data collection (#20) and discussion of limitations (#21).

Research design (criterion #10)

All the evaluations clearly presented the name of the intervention (Q10.1) and presented evaluation questions regarding its effectiveness (Q10.6), either in main text or in their evaluation matrix (Criterion #16), together with clearly defined outcomes (10.3). One example of clear outcome (indicator) definition is given in Table 7. The interventions were often clearly described (Q10.2a and 10.2b) in their findings sections as part of their achievement (that the planned activities were carried out). Evaluation questions regarding effectiveness were usually clearly presented (Q10.6). But for over half of the evaluations the programme timeline was not clearly stated. An important factor for identifying the causal relationship between what was done and what was achieved (temporal precedence) is that the changes should happen after the intervention. At the grassroots and country levels, there were also "missing beginnings", relating to lack of clarity on the programme timeline and the programme components, a point we return to below. Furthermore, few studies clearly described intervention context and programme participants (Q10.4 and 10.5). Some studies that included a section entitled "programme context", but the description there was mainly to give rationale for the intervention rather than present external or cultural factors that might have influenced the effectiveness of the programme, which is our focus. While one evaluation clearly stated that the programme targeted "lesbian, gay, bisexual and transgender (LGBT) people [including men who have sex with men (MSM) as well as intersex and queer], people who use drugs and sex workers", with a list of key populations by programme area, participants were not always clearly described (Q10.5).

Table 7 Example of clear outcome definitions

Outcome indicator	Terms & Definitions	Data collection method(s)
# of advocacy initiatives carried out by CSOs, for, by or with their membership/constituency	Advocacy initiative: An advocacy initiative entails the following activities: · Influencing: Activities aimed at influencing government authorities and power holders. This includes advising, pressuring and persuading state/government officials, private sector representatives, societal actors, multi-stakeholder platforms and the wider public to address the issues / claims of excluded or marginalised groups. · Mobilisation: Activities aimed at creating networks and collaboration to mobilise support necessary for collective advocacy. · Awareness raising: Activities aimed at informing/educating citizens, interest groups and other CSOs on issues/claims of excluded or marginalised groups.	Regular monitoring. Outcome Harvesting
Participation (and satisfaction) in governance processes (political decision-making, mediation and dialogue) by representation of various groups, with special reference to women and youth	Participation: Women and youth are taking part in decision-making processes by speaking up, and negotiating their interests in relation to specific issues during dialogues in formal or informal spaces. Satisfaction: The extent to which women and youth are satisfied with formal and informal participation in decision-making processes.	Participation Regular monitoring KII interviews Outcome Harvesting Satisfaction Survey/Focus group discussions This indicator will only be measured during MTR and evaluation

Source: Evaluation of CARE Nederland's 2016-2020 Programme Every Voice Counts – Inclusive Governance in Fragile States (Aaron, 2021).

Methods to evaluate effectiveness (criterion #11)

Causal claims were clearly stated in the findings sections in the majority of the evaluations (Q11.1). One example of a clear causal claim was: "Creating partnerships with government and other CSOs and NGOs has been pivotal to achieving successful outcomes... The case studies provide an analysis of success factors and lessons from community-led responses and advocacy work at national, regional and global level that have contributed to improved health and rights outcomes for people from key population communities... [and] the programme has made a significant contribution towards improved sexual and reproductive health and rights, fulfilment of human rights and strengthened capacity of key populations' organisations and networks in the countries where it works" (Rainforest Alliance Programme Evaluation, INTRAC). All the reports adequately addressed whether the effects on the outcomes were directly observed by the evaluators (Q11.2). For example, in some programmes, the change in outcomes was observed through the key performance indicators in the programme's monitoring system. In all cases the evaluators observed the effect on the outcome variables (Q11.2), for example: "peer leaders from Bekasi, interviewed for the evaluation, reported positive changes in ten health centres they monitor..." (Bridging the Gaps programme evaluation).

Nearly half of the SRHR programme evaluations used a quasi-experimental design, which observed changes in outcome relative to a comparison group (Q11.3), and examination of alternative causal hypotheses was done in half of the evaluations (Q.11.6). One case that used CA clearly stated for each case selected for contribution analysis that factors outside of the intervention were considered. Another presented lists of the roles of external driving and constraining factors, such as the international context and the effects of the COVID19 pandemic. The majority of evaluations showed an overall timeline (e.g., "the programme was implemented between 2016 to 2019"), without a timeline for the individual activities, but not all of the reports clearly reported the timeline showing that implementation of the intervention preceded the observed changes in outcomes (Q11.4). However, causal mechanisms linking interventions and outcomes were clearly described only in half of the reports (Q11.5).

For D&D, where the evaluation often used a more participatory method like OH, where initial outcomes harvests were usually done by programmes staff, few studies clearly described the timeline showing that implementation of the intervention preceded the observed change in outcome (Q11.4). These "missing beginnings" were particularly noticeable at the country and grassroots levels. Causal mechanisms linking interventions and outcomes were not always clearly described in these cases (Q11.5). Rather, many reports included statements like "the change was observed after implementation of the intervention", followed by concrete examples of changes in outcomes. Although this sort of statement is suggestive of effectiveness, it is insufficient to confirm causality. The minimum requirement for articulating cause and effect is to specify what was done, to or for whom, with what observed outcome. Therefore, studies that presented plausible posited causal mechanism did so by articulating it through the ToC (i.e., specific inputs, activities, outputs and outcomes). Yet examination of alternative hypotheses was rarely done (Q.11.6), as "rigorous causal identification" seemed to be frequently considered beyond the scope of these evaluations.

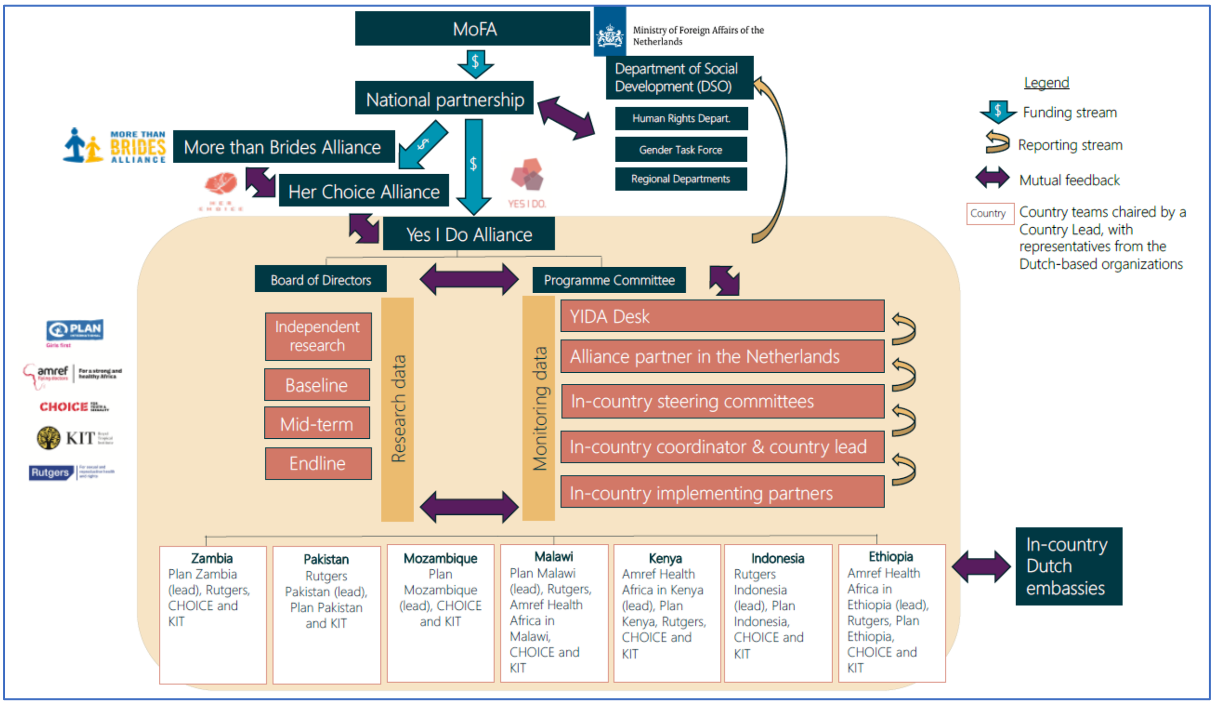
Theory-based approach (criterion #13)

As the programme ToC is a crucial component of the method used in most evaluations, the outcomes analysed were usually justified with reference to it (Q13.5). The ToC was usually presented in the evaluation report (three evaluations referred only to an external source, the programme organisation's website or a mid-term report). Unfortunately, the ToC presented in most evaluations was usually missing some crucial information, such as underlying assumptions, project participants and contextual and external factors (Q 13.1). Furthermore, although the ToC (and log-frames where these were given) often incorporated a list of outcomes, measurable indicators linked to these outcomes were not always provided for D&D programme evaluations (Q13.2). An example of a measurable indicator was the existence of inclusive policy and law-making processes measured under the outcome "inclusion of voice of [female] smallholders". Measurable outcome indicators were more likely to be given in SRHR programme evaluations, particularly for community-level outcomes, such as the percentage of girls who married before 18 years old, or services accessed, such as the percentage of girls who used SRHR services. Potential unintended outcomes (e.g., spillover effects) were not presented in ToCs, while in some studies unintended outcomes, including negative effects, that were observed were reported in the findings sections without linking them to the programme ToC (Q 13.4).

Sampling (criterion #14)

The lists of interviewees and documents reviewed were often included as annexes in D&D programme evaluations (Q14.2 and 14.3). Accordingly, the sample characteristics for interviewees were often adequately reported, as these details were usually included in the list of interviewees (Q14.6). For SRHR, the list of interviewees was rarely included, presumably due to confidentiality requirements, particularly important for these programmes, but many did present detailed sample characteristics, including sample size, location, gender and age group. However, stakeholder maps were not given in any programme evaluations (Q14.1), except in one case of SRHR and two D&D programmes (Figure 5), leading to the potential for 'omitted informant bias'. The sampling strategy, sampling process and its justification were clearly given in SRHR programme evaluations, especially in quasi-experimental designs. Weaknesses were noted where the evaluation used OH and only selection of countries was justified, and the selection process of interview participants was not clearly described. Sample selection processes were not adequately described and rarely justified for D&D programme evaluations (Q14.4 and 14.5). For example, one report presented only how country case studies were selected, but not how interview participants were recruited. As a result, it was not possible to judge appropriateness of the sampling strategy in more than half of the evaluations (Q14.7).

Figure 5 Stakeholder analysis

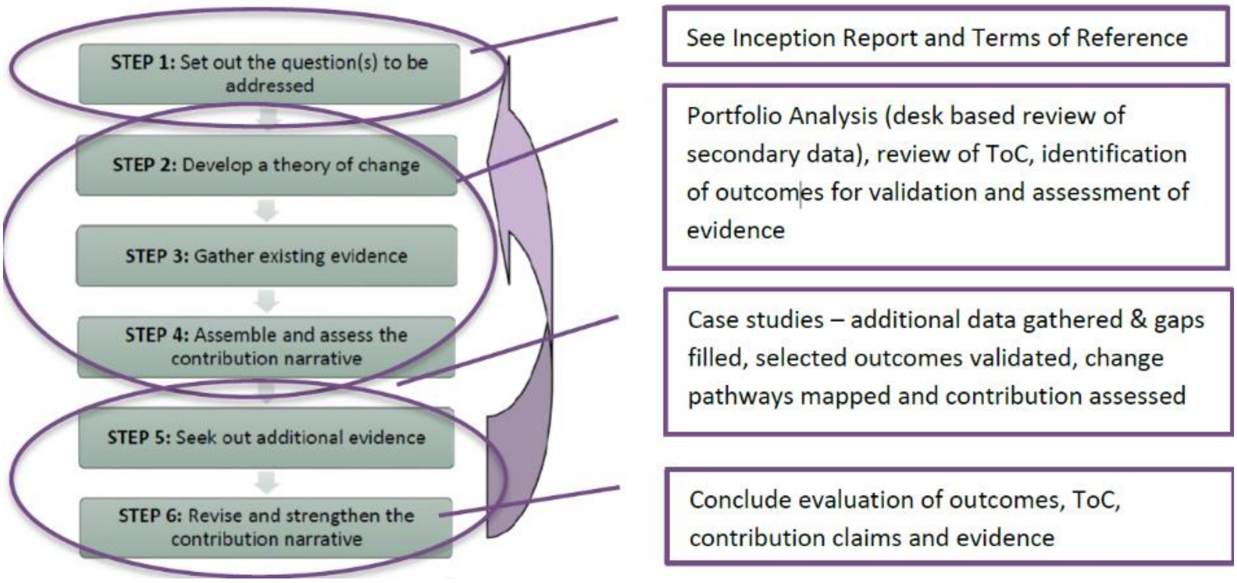

MoFA
Ministry of Foreign Affairs of the Netherlands
$
National partnership
Department of Social Development (DSO)
Human Rights Depart.
Gender Task Force
Regional Departments
More than Brides Alliance
Her Choice Alliance
Yes I Do Alliance
$
$
Board of Directors
Programme Committee
Independent research
Baseline
Mid-term
Endline
Research data
Monitoring data
YIDA Desk
Alliance partner in the Netherlands
In-country steering committees
In-country coordinator & country lead
In-country implementing partners
Zambia
Plan Zambia (lead), Rutgers, CHOICE and KIT
Pakistan
Rutgers Pakistan (lead), Plan Pakistan and KIT
Mozambique
Plan Mozambique (lead), CHOICE and KIT
Malawi
Plan Malawi (lead), Rutgers, Amref Health Africa in Malawi, CHOICE and KIT
Kenya
Amref Health Africa in Kenya (lead), Plan Kenya, Rutgers, CHOICE and KIT
Indonesia
Rutgers Indonesia (lead), Plan Indonesia, CHOICE and KIT
Ethiopia
Amref Health Africa in Ethiopia (lead), Rutgers, Plan Ethiopia, CHOICE and KIT
In-country Dutch embassies
Legend
Funding stream
Reporting stream
Mutual feedback
Country
Country teams chaired by a Country Lead, with representatives from the Dutch-based organizations

Source: "Yes I Do Alliance" programme evaluation.

Methods of analysis (criterion #15)

Regarding whether the evaluations were conducted appropriately, deviations from the standard method were tolerated where the evaluators needed to tailor it depending on the programme contexts and conditions (e.g., restrictions on travel imposed by the COVID19 pandemic). The analysis process was mostly described in detail for SRHR programme evaluations, one good example coming from an evaluation that used outcome harvesting and contribution analysis, which presented a full methodology section including the harvesting and substantiation processes with a clear timeline (Box 2). However, in more than half of the D&D evaluations, it was simply unclear what concrete steps were taken, and therefore whether the evaluation was appropriately conducted (Q15.2-Q15.5). For example, in the case of OH, the criteria for outcomes that were reported was often not clearly described, or the methodology only referred to an external resource (a website from a different organisation). While, in general, the description of data collection was clearly presented, the data analysis process was not clearly given (Q15.6). An example of clear description of the methods used is given in Figure 6.

Figure 6 Presentation of the methods of analysis


STEP 1: Set out the question(s) to be addressed
STEP 2: Develop a theory of change
STEP 3: Gather existing evidence
STEP 4: Assemble and assess the contribution narrative
STEP 5: Seek out additional evidence
STEP 6: Revise and strengthen the contribution narrative
See Inception Report and Terms of Reference
Portfolio Analysis (desk based review of secondary data), review of ToC, identification of outcomes for validation and assessment of evidence
Case studies – additional data gathered & gaps filled, selected outcomes validated, change pathways mapped and contribution assessed
Conclude evaluation of outcomes, ToC, contribution claims and evidence

Source: Final Evaluation of the Sector Partnerships Programme Rainforest Alliance (Allen et al., 2020).

Presentation of an evaluation matrix (criterion #16)

An evaluation matrix is useful as it shows clearly how the evaluation questions are linked to the methods and approaches taken to data collection and analysis. Most studies presented an evaluation matrix (Q16.1) including evaluation questions and data sources, but few included data analysis approach linked to each evaluation question. An example of an evaluation matrix that included most of the essential information looked like Table 8.

Table 8 Example evaluation matrix

Evaluation question	*Approach*	*Data collection methods and sources*
1.1. To what extent have the SPDD programme interventions been effective in making progress towards its three outcome level objectives?	- Inventory of SPDD deliverables from annual reports and interviews with programme staff. - Comparative analysis of actual versus planned deliverables (= specifically for accountability purpose and to assess effects of partner modality). - Inventory of documented progress towards outcome level results, relying on existing OH evaluation and monitoring data. - Substantiation of a representative sample of harvested outcomes (distributed over all objectives, at least 12 outcomes). - Linking the latest outcome harvest results to baseline results, earlier rounds of OH, and the early signs as reported during the MTR.	-Review of existing baselines, plans and M&E documentation as specified in ToR, particularly related to 2018 and 2019. -Grouped interviews with relevant NIMD staff – Programme Managers and Executive Directors of partner organisations/country offices -Interviews with NIMD management, NIMD partners and stakeholders in SDPP programme countries
1.2. What can be said about the plausibility of the contribution relationship between programme interventions and outcomes that have been reported? 1.3.Which of the programme interventions appear to be particularly effective in contributing to programme outcomes?	- Review outcomes harvested by SPDD programme. Select a representative sample of outcomes from latest round of OH results for contribution cases using 'light' CA. -CA for each of the cases by identifying and assessing the contributions categorised as much as possible as capability I, opportunity (O), and motivation (M). This includes further defining specifics of C/O/M contributions, including whether they are internal or external. Consider in particular factors like: changes in democratic space and partner modality. -Assessment of the programme interventions – internal facto–s - to conclude on their relative effectiveness in contributing to programme outcomes.	-Review of existing planning and M&E documentation, in particular documented OH results. - Consultation with NIMD M&E staff /internal steering committee -Interviews with NIMD programme staff and partner organisations (in-country) relevant as informants on selected cases

Evaluation question	*Approach*	*Data collection methods and sources*
1.4. To what extent have the int. lobby and advocacy (L&A) interventions been effective in contributing to outcome level results, and how have these been linked to the implementation of country programmes?	- Inventory of evidence of use of international L&A results in country programmes. -Inventory of main outcome level results with claims of having links with international L&A. -Inclusion of at least 3 of such outcomes in substantiation sample and, if possible, (at least one) in case studies for light CA.	-Document review of relevant planning and M&E document -Interviews with relevant NIMD programme staff and international partner organisations. -Consultation with NIMD M&E staff /internal steering committee

Source: NIMD Strategic Partnership Dialogue and Dissent programme evaluation (MDF Training and Consultancy, 2020).

Sources of information (criterion #17)

The evaluations generally used different types of information sources such as documents, interviews, workshops and focus group discussions, to triangulate the outcomes observed (Q17.1). The list of interviewees often suggested that appropriate sources (e.g., different stakeholders, such as participants, implementers, programme managers, country partners, funders, boundary partners, beneficiaries) were sought (Q17.4). Yet more than half of the D&D programme evaluations did not attempt to guard against subjective selection of cases (Q17.3) as they did not include information about respondent selection, and in the case of document review, data collection coding frameworks were not presented. However, attempts to subjectively select country/project cases were seen in a number of evaluations: one example is given in the Box 3. Few evaluations, of either D&D or SRHR programmes, clearly indicated that relevant sources external to the intervention, such as non-participants, organisations who may have experienced another intervention, or those not targeted by interventions like trade unionists, were consulted (Q17.5). Discussion of issues around recruitment was rarely done, so it was also not clear whether (and why, if any) some people selected for sample collection chose not to take part (Q17.6). In contrast, for SRHR programme evaluations, the data collection more clearly attempted to mitigate cherry picking of cases, either through random sampling (in the case of quasi-experiments) or through purposive selection from diverse group to reflect different points of view. Half of the SRHR evaluations discussed recruitment issues; one study, for example, mentioned security issues in the programme countries as a reason why data could not be collected from some areas (Q17.6). Assessing the appropriateness of data source was often impossible, as relevant information was not found.

Box 3 Example of subjective case selection strategy

"Sampling may be two-stage, if necessary. In the first stage of sampling, subthemes will be selected by the evaluation team, in collaboration with the internal taskforce and in close consultation with the thematic units. The sampling criteria for this stage will be:

- One subtheme per ToC (3 total);
- Clear knowledge gap / avoid duplication of evaluation efforts on the same project / learning priority: limited or no existing evaluations in the same area
- The subtheme constitutes a meaningful part (significant number of projects and funds) of the total ToC budget and scope.
- At least one of the subthemes selected should include the work of the Centre for Research on Multinational Corporations (SOMO).

A second stage of sampling will be conducted only if the selected subthemes encompass a large number of projects and/or outcomes. In this case elements within the selected subthemes will be sampled:

- The sampled projects should work significantly on that subtheme and not only have a minor contribution to the subtheme (a budget threshold is not set, however at least 2 projects should be included)
- Lack of existing evaluations of the same projects (for instance, not included in the MTR)
- Unintended outcomes (both failures and success) were experienced by the projects"

Source: End-term evaluation of the Strategic Partnership between Oxfam Novib and SOMO 'Towards a Worldwide Influencing Network' (2016-2020) (Arkesteijn et al., 2021).

Triangulation (criterion #18)

In most cases, triangulation was incorporated as an essential part of the evaluation design. For example, for the method most commonly adopted, OH, outcomes harvested by programmes teams were subject to validation or substantiation through further interviews by the evaluators. Usually, therefore, this was data triangulation, using data from different locations, times and participants, but "methodological triangulation" was also done in some cases, for example by combining two evaluation approaches (e.g., OH with CA or quasi-experimental design with qualitative data on processes). Two evaluations used investigator triangulation: in the case of the "Towards a Worldwide Influencing Network" programme evaluation, more than one interviewer was used to collect the data; the report on the "Jeune S3" programme noted that "During the implementation of the evaluation, room for exchange of different perspectives amongst the evaluation team members was provided. (Annex, p.22)" Box 4 presents an example of a triangulation approach that is clearly explained.

Box 4 Example of triangulation and evidence rating

"Guideline for rating the level of evidence, mainly to be based on the degree of triangulation.

Strong: data on both change and contribution verified through one or more credible external data sources, in addition to internal sources. Divergent perspectives and alternative contributions explored.

Medium: data on change and/or contribution verified through one or more credible external data sources, in addition to internal sources, but data gaps still remain.

Weak: data on change and contribution verified through internal data sources only."

Source: Bridging the Gaps End Evaluation (Napier et al., 2020).

Discussion of bias (criterion #19)

Controlling for and discussing bias seemed to be the most challenging criterion to meet, which was reflected in the coding. Feedback on the preliminary coding from NGOs and evaluators revealed that they did take these potential biases into account and attempted to mitigate them, but this was not usually mentioned in the reports. Some evaluations mentioned a risk of bias in the limitations sections, but measures taken to guard against them were not adequately reported. The argument made for this choice was usually that the focus was not on identifying a causal relationship but in learning from good and bad practices. Data triangulation was often used as a way to mitigate against bias. However, it is unlikely that data triangulation can always deal with any sort of bias. For example, interviews often started with questions about the programme, and then went on to ask about achievements and possible causal claims, which is a clear form of "anchoring bias". Only one evaluation clearly stated that conflict of interest was considered (Q19.8): "The entire evaluation team including national experts held no stake in the programme and was therefore unbiased. In contrast, it is possible that interviewees displayed a positive bias in their answers in case they were hoping to receive renewed funding as a result of a positive evaluation. (Jeune S3 Programme; p.11)" In Box 5, we present the main sources of bias in turn.

Box 5 Sources of bias in impact evaluations

Alternative causal explanation (19.1): some evaluations presented implementation issues and contextual factors that might have affected the outcomes. For example, the Green and Inclusive Energy Evaluation (IIED and Hivos) linked policy engagement outcome achievements with effective lobby and advocacy strategies implemented under the D&D programme, as well as external enabling factors, including decentralisation in the energy sector which provided opportunities to provide local technical support, and general concerns in the international community and general public about climate change and enabled agenda setting by the CSO partners. The evaluation also linked desired policy changes around harmful fossil fuel subsidies, which were not achieved, to external political and civil changes, the discovery of fossil fuels that Dutch embassies supported, and internal factors such as the collaborative advocacy approach that was adopted which made it harder to criticise fossil fuel companies openly. One evaluation noted that "the success...appears to be the lobby power of the CSOs itself, rather than new capacities installed through training... [or that] the success factors here were to use existing systems, rather than prior training." Another acknowledged that "other external factors are important in this (success) story". Apart from this, no clear statements on alternative causal explanations were found. One study stated that, using contribution analysis principles, they explored the potential contribution of other actors, but discussion of these factors was not done. As a result, we observed that the attempt to rule out competing causal explanations or contributory factors was not done in any of the evaluations using ‘small n’ approaches.

Evaluator bias (Q19.3, 19.4 and 19.7): biases caused by evaluators’ own positions and assumptions, were very rarely mentioned. One type of bias we thought might be discussed related to “confirmation bias”, which can be mitigated, for example, by recording interviews and comparing notes by multiple interviewers. Only a few studies suggested that this sort of mitigation measure was taken. One evaluation indicated coders were blinded during the outcome mapping session: “The INTRAC team mapped each outcome against the Theory of Change; this was‘done 'blind’ to avoid bias by similar coding completed earlier by Aidsfonds’ M&E team” (Bridging the Gaps programme evaluation; p.71). Two others attempted to mitigate evaluator bias through investigator triangulation, although less clarity was given on whether and how the investigators’ notes were compared.

Respondent bias (Q19.5 and 19.6): “courtesy bias” or “political correctness bias” are forms of social desirability bias, “positional bias” includes errors of attribution, and “self-serving bias” or “self-importance bias” concerns positioning oneself or one’s organisation at the centre of events. These biases can be mitigated by, for instance, drawing up interview schedules to avoid leading questions, or blinding participants to the purpose of the evaluation (i.e., not mentioning the intervention, at least early on in interviews). Only a few studies mentioned this bias. To mitigate respondent bias, one evaluation incorporated ‘Social Presencing Theatres’ in focus group discussions, “to elicit unbiased answers since humans are more accustomed to modifying their words than their gestures in accordance to outside expectations” (Her Choice Programme Evaluation). Another used what they called a ‘double blind’ methodology, meaning that neither the researcher nor interviewees as part of the review knew who the client was, although it should be noted this is not the same as masking of knowledge about participation in the intervention. One-third of the evaluations attempted to protect against “recall bias”. For example, outcome harvesting was started in the final year of implementation in one programme evaluation. While it was usually clear when the evaluation was conducted, in many cases it was unclear when the data were collected with respect to the interventions and outcomes.

Source: authors.

Data collection (criterion #20)

The processes for collecting data were clearly presented in most of the evaluations, indicating how, when and with whom interviews, workshops and FGDs were conducted and recorded. Often, questionnaires and/or interview protocols were presented (Q20.3). Less clear was whether data codes, categories or themes were structured around the ToC (Q20.1). Regardless of the evaluation methodology, results and findings sections and analysis protocols were rarely clearly linked to the ToCs, although where OH was used, it can be inferred that each outcome category observed was referred to the ToC by the nature of the methodology. Few evaluations linked their data collection protocols to possible alternative hypotheses (Q20.2). As indicated above, document review was conducted in most evaluations, and the list of documents given, but it was unclear how data were collected from them (i.e., what data collection codes were used) (Q20.4).

Discussion of limitations (criterion #21)

When evaluation questions were clearly stated, the findings sections generally addressed all of them. Most reports included a summary section that clarified the link between the findings and evaluation questions (21.1). Similarly, the implications or recommendations were clearly linked to the findings, by virtue of the fact that this section usually came immediately afterwards, hence the link was clearly made (21.3). Almost all of the evaluations had a “limitations” section (21.2) where the evaluators explored various challenges from risk of bias to limited data availability due to the pandemic, with different depths of exploration. Limitations due to data or resource availability as a result of the COVID19 pandemic were usually mentioned. Some evaluations attempted to mitigate the resulting biases. Most SRHR evaluations clarified in their main text that their research complied with ethics (anonymity, informed consent and confidentiality). But, for nearly 20 D&D programme evaluations, we could not confirm that the research complied with ethical standards (anonymity, informed consent and confidentiality). There is a stronger tradition in evaluation in the health sector to address ethical issues regarding the collection and use of personal information. However, feedback on the preliminary assessment from the NGOs and evaluators revealed that ethical issues were considered, but not reported.

Chapter 4 What are the implications of the findings?

In this chapter we return to and aim to answer the five evaluation questions from chapter 1. We discuss the findings in light of other approaches used in the evaluation of lobby and advocacy, presenting implications of the study for the design, conduct and reporting of evaluations of lobby and advocacy programmes. This discussion is relevant for 'small n' impact evaluation more generally.

4.1 Discussion of findings in relation to evaluation questions

EQ1: What evaluation methodologies have been used and were they adequately applied in practice?

The most frequently used methods to determine the effectiveness of the programmes (the causal effect of the interventions on the outcomes) were Outcome Harvesting or, sometimes alongside, Contribution Analysis. But a substantial proportion of the evaluations did not specify any method used to articulate the contribute of the programme. This is not the same thing as saying that they did not specify data collection approaches (like FGDs and document review), or sampling procedures, for example – they often did. Articulating the approach to collecting data is necessary but insufficient for determining how the evaluation question is to be addressed.

Where a method for evaluation effectiveness was used, although not in all cases, the methods undertaken were those that the evaluators had envisaged and appeared to be implemented appropriately. Frequently, OH was used whereby the programmes organisations had built the approach into internal monitoring, evaluation and learning (MEL) systems, providing a participatory approach to building capacity in results-based management. Drawing on these initial harvests, the evaluators would then conduct a substantiation process, where credible outcomes were triangulated with information from other sources. This was undertaken in some, but not all, causes. Sometimes, although more rarely, this approach was combined with another method like CA, MSC or MAPP.

The evaluation reports were often long enough for information on the evaluation methods, data collection and analysis to be reported transparently, either in the main text or appendixes. However, the standards of reporting were often inadequate for our assessment, and this lack of clarity was also reflected in our assessments of the causal claims that were being made.

EQ2: Are the evaluation methodologies consistent with the updated IOB evaluation quality criteria?

As noted above, the updated IOB evaluation quality criteria were not available at the time the evaluations were designed, hence this assessment is not a performance review of the evaluations we reviewed. Our summary assessment of the evaluations by IOB Criteria is below.

Research design (Criterion #10): evaluation questions regarding effectiveness were usually clearly presented. But for over half of the evaluations the programme timeline was not clearly stated.

Methods to evaluate effectiveness (Criterion #11): an important factor for identifying the causal relationship between what was done and what was achieved (temporal precedence) is that the changes should happen after the intervention. Few studies clearly described the timeline showing that implementation of the intervention preceded the observed change in outcome. At the grassroots and country levels, there were often "missing beginnings", relating to lack of clarity on the programme timeline and the programme components. It was also not usually clear what the relative contribution of the programme was to the outcomes observed, or what were the causal pathways or mechanisms of change.

Theory of change (Criterion #13): a programme theory was usually presented which incorporated a list of outcomes, but measurable indicators linked to these outcomes were not always provided, especially for D&D programme evaluations.

Sampling (Criterion #14): the sample characteristics for interviewees were often adequately reported, but stakeholder maps were given in only one evaluation, leading to the potential for "omitted informant bias".

Analysis (Criterion #15): the analysis process was mostly described in detail for SRHR programme evaluations. However, in more than half of the D&D evaluations, what concrete steps were taken, and therefore whether the evaluation conduct was appropriate, was unclear.

Evaluation matrix (Criterion #16): most studies presented an evaluation matrix, but few included the data analysis approach linked to each evaluation question.

Information sources (Criterion #17): the list of interviewees often suggested that appropriate internal sources were sought. Few evaluations clearly indicated that relevant sources external to the intervention, such as organisations who may have experienced another intervention, or those like trade unionists not targeted by interventions, were consulted. Recruitment problems were rarely discussed, so it was also not clear whether (and why, if any) some people selected for sample collection chose not to participate.

Triangulation (Criterion #18): in most cases, data triangulation was incorporated as an essential part of the evaluation design. But investigator triangulation, where multiple investigators compared notes after interviews, was rarely reported as having been implemented.

Bias (Criterion #19): few evaluations attempted to rule out important sources of bias that affect any causal study: namely alternative causal claims, contributory factors, and predictable respondent or evaluator biases. Data triangulation was often used as a way to mitigate against bias. However, it is unlikely that data triangulation can always deal with any sort of bias. For example, interview protocols often began with questions about the programme, before asking about achievements and possible causal claims, which is a clear form of "anchoring bias".

Reporting of data collection and analysis (Criterion #20): the processes for collecting data were clearly presented in most of the evaluations, indicating how, when and with whom interviews, workshops and FGDs were conducted and recorded. But they were often not clearly linked to programme ToC.

Limitations (Criterion #21): the evaluations usually discussed limitations due to data or resource availability as a result of the COVID19 pandemic. A few evaluations mentioned possible sources of bias, but most did not attempt to mitigate the resulting biases or see it as a key role of the study.

EQ3: What are the common characteristics of appropriate methods for evaluating the effectiveness of L&A?

Existing methods papers agree that it is particularly challenging to evaluate the effectiveness of L&A. Van Wessel (2018) argued that the disadvantage of 'small n' approaches like Most Significant Change and Outcome Harvesting are that they are highly intervention-focused, and pay scant attention to the role of the context in affecting change ("self-serving" or "intervention-centric bias"). This suggests a strong understanding of the context is an important prerequisite for evaluating causal claims. She also argues that an appreciation of systems dynamics including non-linearity (e.g., interaction effects, feedback loops) and the need to understand that L&A strategy planning is only partly plannable and necessarily adaptive. Both van Wessel (2018) and Teles and Schmitt (2011) thus stressed the importance of expert skills in making evaluative judgements rather than methods. While we are not discounting the importance of expertise of the evaluation team in the topic area – team composition and team expertise being key factors influencing the quality of the evaluation – the rationale for this project is to aim to identify reliable methods for evaluations of support to L&A and to improve the design, conducting and reporting of credible evaluation approaches.

Barret et al. (2016) emphasised the role of theory of change and present outcome indicators for L&A, and suggest that it is useful to limit the number of outcomes being evaluated, so as to balance accuracy of the approach with resources available for evaluation. We adopted a ToC-based approach in this study, going beyond Barret et al. to discuss the basis for the causal claims in the ToCs.

The minimum condition for addressing a question about effectiveness, meaning the causal effect on, or contribution of the programme to, the defined outcome(s), is that it is clear when and for whom the intervention was undertaken, what outcomes were achieved, and what were the likely causal pathways (intermediate outcomes) and contextual factors that might provide competing possible explanations or contributory factors. This implies, firstly, that evaluations should be based on, and reported around, a programme theory of change. A good example of presentation of and use of ToC can be found in the Hivos' report on Open Up Contracting Program. It clearly outlines outputs, intermediate and final intended outcomes, along with indicators, assumptions and intervention logics that link these elements. Participants and project-affected people were explicated in each element of outputs and outcomes, while explicit descriptions of contextual or external factors were mostly found in the outcome analysis section rather than the ToC section, a tendency observed in many other evaluations. Theory of change is a crucial step of many 'small n' approaches, and is important more generally in programme evaluation. For example, a clear advantage of articulating the ToC is to avoid the problem of "premature impact evaluations", where data are collected and analysed before changes in outcomes can be realised. Indeed, the Green and Inclusive Energy Evaluation (IIED and Hivos) evaluation also noted that "five years is a short period of time to achieve the long-term institutional changes as formulated in the TOC, especially since they not only refer to policy change but also implementation", a point which is applicable to (and noted by) evaluations of policy change and implementation more generally.

Secondly, it implies that some method is needed to substantiate causal claims being made and to articulate the likely contribution of the programme activities to the outputs and outcomes achieved. These conditions are most likely to be met for Type I methods such as Contribution Analysis and Process Tracing. A Good example of conduct and presentation of substantiation as part of contribution analysis can be found in the annex of the NIMD Strategic Partnership Dialogue and Dissent programme final report (Zuijderduijn et al. 2020), which clearly showed the reliability of data on contributing factors was rated based on how well it was triangulated (by whom, and by how many data sources).

Anguko (2019) argued that, where they are possible, evaluations that use a combination of 'small n' mechanism-based approaches and 'large n' counterfactual analysis are able to analyse causal claims and provide estimates of effect magnitudes. However, the application of appropriate qualitative causal inference approaches can clearly be strengthened. This review found that few evaluations properly applied a 'small n' causal inference approach that systematically unpacked and assessed causal mechanisms, and resulted in substantiated causal pathways vested in an explanation of how an intervention is leading to a change in a particular context (i.e., taking into consideration other causal factors). Gardner and Brindis (2017) argue for greater use of Contribution Analysis and similar approaches in advocacy and policy change evaluation to analyse systems change, together with experimental and quasi-experimental approaches to assess changes in quality-of-life outcomes among target populations, but note they may not be appropriate methods to link to advocacy efforts themselves.

Several of the SRHR programme evaluations used a combined approach, which was possible as the objectives of the sexual and reproductive health programme evaluations usually extended to service delivery and health outcomes, where there were sufficient numbers of recipients to use statistical methods. In these cases, the mechanisms for affecting capacity, support to L&A, policy engagement and policy change were assessed using methods like OH and CA, while the effects on service delivery and health outcomes were assessed using quasi-experimental methods. No D&D programmes used a combined approach. Although the endpoint outcomes for D&D programmes tended to be 'small n' in nature, it may have been possible to evaluate the effectiveness of L&A support for capacity building, where there may have been sufficient numbers of CSOs or CSO staff members for 'large n' approaches.

Where projects work in multiple global regions and countries, with multiple partners, a suitable appropriate approach is to provide a descriptive overview of the portfolio at global or regional level, and evaluate effectiveness in a small sample of cases (countries, projects, or L&A trajectories) chosen according to some transparent selection process. This approach was adopted by many of the programme evaluations, although the justification for the cases chosen was usually unclear.

EQ4: What were the common characteristics for the less suitable methods?

Drawing on the evaluations that were assessed as being at 'low confidence', we present a synthetic example of an approach that would not in our view be able to address contribution or attribution of programmatic support to L&A:

- The evaluation did not collect data on outcomes that were intended in the theory of change but not achieved.
- The evaluation did not specify a method that was used to assess causal claims. For example, where a case study approach was used to collect and analyse data, it did not articulate relevant causal pathways using the ToC.

- Where the approach did specify a method used to evaluate effectiveness, the outcomes harvested were not independently verified by the evaluators through data triangulation.
- Outcomes were collected from those who were part of the programme or who participated in the L&A activities, but not from informants that did not participate in the programme.
- There were 'missing beginnings', so there was very limited analysis of activities, especially at the grassroots level, to demonstrate that the L&A actions undertaken by CSOs were related to the actions of the programme itself.
- There were weak measures of outcomes, such as on capacity building, or, where outcome measures were potentially strong, 'missing middles' that demonstrated feasible causal pathways to their achievements from the activities undertaken.
- Predictable biases were not avoided. For example, interviews were poorly designed, commencing with questions about the programme of interest, and proceeding to ask about possible outcomes, a clear example of anchoring bias.
- No attempts were made to analyse alternative causal pathways that may have contributed to the outcomes being achieved (or not), whether other interventions that were occurring at the same time, or contextual factors.

Regarding data collection we note that, for some outcomes, the measure was relatively strong, in that it was objectively verifiable (e.g., a policy change), but the contribution story less credible owing to the length of the causal pathway. In other cases, the contribution story was potentially strong but the measures of change were often weak (e.g., capacity building), although good practice examples are reported above. But there is also the issue of "probitive value" when selecting informants, where the choice of respondent is related to the quality of the information provided. That is, not all data sources are of equal value. It was rarely clear how the choice was made about whom to sample among those participating in the programmes (usually grassroots CSO and programmes staff) and those that did not participate who might have had different, but informed, perspectives about the programme's achievements.

EQ5: What can be said about the achieved results of the supported partnerships?

We extracted middle-level theories for D&D and SRHR drawing on evidence provided in studies that were assessed as being of medium and high confidence (Chapter 2). These theories articulated how the L&A interventions acted to change knowledge, attitudes and behaviours of targeted groups, including CSOs, government, private sector and community leaders and members. The MLTs then synthesised evidence on the factors that enabled and derailed the achievements. Regarding our confidence in the outcomes achieved, the evaluations rarely provided a clear explanation of the contribution of the programme activities to the outcome and the strength of the evidence, and where these were reported, they were often rated as 'medium' or 'strong'. Hence, for most of the outcomes harvested, the contribution of the L&A support was unclear.

However, we also noted that there appeared to be clear bias in the evaluations on the reporting and analysis of positive changes. For D&D programmes, 89 percent of the outcomes reported were positive changes (improvements), and for SRHR 80 percent were positive changes. These were measured in the areas of capacity development, support to lobby and advocacy efforts, policy engagement, policy change, empowerment and access to SRHR services. While not discounting the effectiveness of the programmes that were evaluated, we would also expect there to be negative and null (no change) outcomes, for most programmes, especially complex ones operating in the area of L&A where only a proportion of activities are expected to yield successful outcomes, due to a high rate of failure expected for any single activity or tactic (Teles and Schmitt, 2011). The high rate of success in the programme evaluations here suggested that negative or null outcomes are simply not being captured and reported by the evaluation methods used.

The underreporting of negative effects or null effects is in part caused by low incentives to delve into negative effects or elucidate null effects; "self-serving bias" or "intervention-centric bias" leading to an overestimation of the role of an intervention vis-à-vis other causal factors; and methods choice. Some methods (e.g., OH, MSC) may enhance a bias toward positive effects reporting, requiring work on the part of the evaluators to explicitly capture alternative causal pathways, or to assess outcomes that were intended but did not occur, discussed below.

4.2 Implications for evaluation design, conduct and reporting of L&A programme evaluations

This meta-evaluation provides a basis for further development of rigorous qualitative impact evaluation methodologies also appropriate for L&A. We presented MLTs for D&D and SRHR drawing on the observed outcomes, which were assessed as being of at least 'moderate confidence'. These articulated key enablers, derailers and safeguards, but the analysis was constrained by the evaluation design, conduct and reporting, as well as the engagement with all potential outcomes, not just those that were achieved.

Generally, the evaluations did notppear to be concerned with, or able to address, sources of bias and the attribution or contribution challenge. This included evaluations based on participatory methods like Outcome Harvesting and those using approaches like Contribution Analysis. Causal pathways and mechanisms of change should be at the core of evaluations of programme effectiveness. An approach that was not explicitly used in any of the evaluations is Process Tracing (PT) (Ford et al., 1989) (Box 6).

Box 6 Description of methodologies discussed in this section

Process tracing (Ford et al., 1989; Vaessen, 2020): a case-based approach that aims to empirically test the causal mechanisms that link program components and outcomes. Major steps involve: (1) formulate hypothesized causal mechanisms for the observed outcome (2) collect the observable data to be used for testing for the presence of the mechanisms (3) Assess the evidence for each hypothesized mechanism using four tests (the straw-in-the-wind test, the hoop test, the smoking gun test, and the doubly decisive test).

Contribution analysis (Mayne, 2020): an approach that aims to compare the intervention's ToC against the evidence, by constructing "contribution story" to demonstrate the contribution made by the intervention, while considering the role played by external/contextual factors on outcomes. Major steps involve (1) set out the cause-effect question(s) (2) draw up a ToC (3) gather existing evidence on the ToC (4) construct "contribution story" outlining whether the intervention was implemented as planned, how other factors influenced, and whether the expected outcomes were achieved (5) strengthen the credibility of the contribution story by searching for additional evidence.

Qualitative Impact Protocol (QuIP) (Copestake et al, 2019): a standardised approach to generating feedback about causes of change, relying on the testimony of a sample of programme participants. Major steps involve: (1) field data collection by two researchers, typically with 24 semi-structured interviews and four focus groups, which has the following characteristics: purposive selection of interviewees, blindfolding where possible and use of pre-formatted data collection spreadsheet. (2) coding by a data analyst (different from field researcher), and semi-automated generation of summary tables and visualisation (3) dialogue and sense-making between researchers, commissioners and other stakeholders.

PT combines elements of Outcome Harvesting and Contribution Analysis to collect and analyse evidence on the activities conducted, the outputs and outcomes achieved, and the contribution of the intervention to the outcomes, including its relative significance to other contributory factors (Patton, 2008). The advantage of the method is that it aims explicitly to articulate and evidence possible causal pathways for change, providing a contribution assessment which allows for the existence of multiple causal pathways and which can be applied retrospectively to processes and outcomes that were not specified at the outset. Another approach which was designed with the intention of addressing biases in 'small n' evaluation is the Qualitative Impact Protocol (QuIP). Copestake et al. (2019) proposed QuIP for 'small n' settings; it has overlapping features with the common qualitative methods like outcome harvesting, but addresses the attribution challenge explicitly by attempting to eliminate bias by focusing data collection on all drivers of change for the outcomes of interest with both respondents and enumerators 'blindfolded'[6] to the intervention being evaluated, thus reducing respondent and evaluator biases and potentially addressing attribution. This approach is evidently possible, especially where evaluations done by consortia that incorporate or are led by local partners. While this is a strong tool in situations where establishing independent

[6] The difference between blinding and blindfolding is that blindfolds can be removed at a later period of the data collection and analysis process (Copestake et al., 2019).

counterfactuals is not realistic, collecting data from a sub-sample of non-participating organisations or groups that were part of or influenced by other programmes, may also help in identifying non-intervention drivers of change (confounders). But even in a non-blindfolded evaluation, anchoring bias can and should be avoided through careful interview design.

Where a method is used like Outcome Harvesting and Most Significant Change, which have the advantage of building in participation in the evaluation from those involved in programming, it is important that the outcome harvested by programmes staff and/or participants are, firstly, substantiated by evaluators through enquiry and data triangulation. Secondly, there is likely to be a need for additional analysis by the evaluators to assess the possibility of alternative causal pathways, in order to address contribution, or to assess outcomes that were part of the programme theory of change, but did not occur. The latter might be achieved through enquiry and analysis of outcomes that were part of (pre-specified by) the programme theory, but not necessarily mentioned in outcomes harvested by programmes staff.

We also present specific implications for evaluation design, conduct and reporting, firstly for evaluators.

- Design: as the evaluations are necessarily theory-based, they should be designed clearly around the programme ToC. The ToC therefore forms the structure for which the evaluation methodology is designed, data are collected and causal claims verified and presented. The approaches used should address biases in the causal claims being made explicitly, including outcomes that were not achieved, alternative causal claims, respondent bias and evaluator bias. It is important to assess CSO capacity in order to evaluate whether competent L&A was work even if it had no effect on endpoint outcomes like policy change and implementation; for example, this might include the capacity to keep campaign material ready until the time is right. We acknowledge that contextual factors such as shrinking civic space influences the options of feasible evaluation methods. However, evaluations that draw closely from the programme theory of change should also be able to engage better with outcomes that were not achieved, and collect data to help understand why that was the case.
- Conduct: in most areas of conduct, the evaluations showed limited engagement with some evaluation quality criteria, because of space and time constraints in their reports, and not necessarily the inappropriateness of their evaluation approaches. The one exception related to the discussion of bias, where it was generally found that respondent bias, evaluator bias and exploration of alternative hypotheses was not addressed. Evaluations that draw more clearly on the programme theory of change, and thus assess both outcomes that were achieved and those which were intended but not achieved, and which use clear methods to address predictable biases (e.g., blindfolding, well-designed interview schedules, engagement with informed outsiders, analysis of competing causal pathways) will be more useful for decision making. It may be possible to incorporate these additional components to evaluations that draw on methodologies like OH or a combination of OH and CA. In the case of large-scale, multi-country or multi-component programmes in L&A, it is not realistic to conduct rigorous evaluations in each case due to time and resource limitations. Selection of a few case studies using subjective (but transparent) selection criteria, and evaluation of these cases using a rigorous theory-based approach is a solution. Detailed examination of contextual factors in each case study can provide valuable learnings for future programme design.
- Reporting: in addition to transparent reporting about the methods and data collection approaches, for which a reporting check-list could be provided for inclusion in the report annexes, it would also be useful for evaluations to provide clearer causal claims about achievements. However, many of the causal claims made in the studies are necessarily theory-based, and there is a potentially long causal pathway between what was done and what was achieved, especially where the achievement was a distal outcome like a policy change or service delivery improvement. As noted above, these activities, outputs, intermediate and final outcomes should be articulated in the programme theory of change on which the evaluation is designed. It would also be useful for evaluations to present the causal claims according to these ToCs, preferably by presenting tabular or flow diagrams that indicate the causal claims being made together with relevant intermediate steps, contributing factors and assumptions. This will also help address the common problem found in the evaluations of "missing beginnings", where it was not clearly reported what was done as part of the intervention at grassroots levels.

The distinction between issues that relate to design and conduct, and those which relate to reporting, are important. We suspect that, in the evaluations we reviewed, some low scores were likely to be due to reporting failures rather than conduct failures in evaluations (e.g., investigator triangulation may have been implemented in a number of cases but was not mentioned in the methodological sections or appendices of the report). But in other cases, such as adequately dealing with respondent bias, low scores were due to be design and conduct failures.

For commissioners, in CSOs or government, we noted that a significant minority of the evaluations reported their data collection methods but did not indicate a method to evaluate effectiveness that could verify causal claims. Without clear evaluation methods, collected data do not help the evaluators with analysis of contribution to outcomes achieved. While appropriate methods vary from one programme to another, depending on its context, budgetary constraints, programme size and sample size (whether 'small n' or 'large n'), further guidance could be provided to evaluators on the methods that can be used to evaluate causal claims for particular types of programmes and outcomes, drawing on IOB's updated evaluation quality criteria. Some studies clearly stated that their evaluations focused on contribution, but others did not, which suggested that, for evaluations with an effectiveness question, guidance would be especially helpful for devising evaluation questions that more clearly address contribution or attribution. Research may be needed to assess whether it is possible to address predictable biases that arise in the data collection and analysis process of methods commonly used to evaluate L&A, such as OH. Research is also needed on appropriate methods that are suitable for evaluating effectiveness in the lobby and advocacy for systemic changes (public and private sector actors' policies and practices).

There is also a need for guidelines, including a reporting template and good examples from past evaluations, so authors of future evaluations can include all the necessary and sufficient details to satisfy commissioners that the evaluation had been conducted and reported appropriately to answer the questions posed about programme effectiveness. Guidelines and checklists are commonly provided in other areas, such as for evidence and gap maps (White et al., 2020). We also believe that the guidance can incorporate or suggest methods (or add-ons to existing methods) which can help provide assurance against predictable biases, including those that are less commonly used, if at all, such as whether blinding (or blindfolding) is possible at any stage in the evaluation. Other approaches, used in other areas of monitoring, evaluation and learning, include specific efforts by commissioners to ensure that there is learning from intended outcomes that were not achieved, such as "failure fests". At a minimum, CSO and government commissioners should indicate that evaluations clearly draw on the theory of change posited by the programme, and aim to collect data on all relevant outcomes, not just those that were successfully achieved, in order to learn from successes and failures more systematically. Examples of key methodological principles to be considered in designing, conducting and reporting qualitative causal inference studies are presented in Table 7.

Table 9 Issues to be considered in checklists for qualitative causal inference

Updated IOB Evaluation Quality Criteria	*Example checklist item*
U 10. The research design is clearly elaborated and shows how the research results will contribute to answers to the evaluation questions	The evaluation presents clearly defined method that is used to examine effectiveness (contribution and/or attribution) of a clearly defined intervention, programme or policy on a clearly defined outcome.
U 11. The methods are appropriate to evaluate effectiveness	An appropriate method is used to assess attribution and/or contribution that can measure and validate effectiveness relative to other programmes and relevant contextual factors.
U 13. The indicators or result areas are appropriate to capture the planned results along the different levels in the ToC	The evaluation addresses possible sources of bias in interventions being reporting or outcomes being collected with reference to an explicit theory of change, and aims to collect evidence on outcomes that were achieved and those which were intended (according to the actions implemented) but not achieved.
U 14. Justified choice of sample, cases and information sources (e.g. choice of countries, projects, organisations and persons)	The evaluation addresses sources of bias in possible causal claims being measured through a sampling strategy that is presented transparently and appropriately justified.
U 15. The analyses are appropriate, given the chosen research design	The method used to assess attribution and/or contribution is conducted appropriately, which may include – for approaches like OH and MSC – additional components of the evaluation

Updated IOB Evaluation Quality Criteria	*Example checklist item*
	that are designed to measure and validate effectiveness relative to other programmes and relevant contextual factors.
U 16. Summary of the methodology in an evaluation matrix	The evaluation presents a matrix linking questions on effectiveness to study design, methods and data collection.
U 17. Sufficient independent information sources	The evaluation addresses sources of bias in causal claims being made through a sampling strategy that includes informed independent information sources to avoid “omitted informant bias”.
U 18. Triangulation of results from different information sources	The evaluation reports the methods of triangulation used, such as data triangulation, methods triangulation or investigator triangulation.
U 19. Discussion of bias	The evaluation addresses sources of bias in causal claims being made including social desirability bias or errors of attribution by participants (which can be addressed through more careful design of interview schedules or ‘blindfolding’), “confirmation bias” and other biases on the part of the evaluators (mitigated through recording interviews and comparing notes by multiple interviewers), and engaging with external factors influencing the observed outcomes.
U 20. Systematic, complete and transparent description of the data collection and analysis	The evaluation reports transparently how the different sources of data were collected and analysed (e.g., interview schedules and document review coding sheets).
U 21. Discussion of the limitations of the evaluation	The evaluation clearly links findings to the data collection and analysis and contains a discussion of limitations of design or conduct.

References

Allen C, Arhin A, Mundzir I, Jain N and Pratt B (2020) Final Evaluation of the Sector Partnerships Programme Rainforest Alliance. INTRAC.

Anguko, A. (2019) Process tracing as a methodology for evaluating small sample size interventions. eVALUation Matters Second Quarter 2019 African Development Bank. https://idev.afdb.org/sites/default/files/documents/files/Process%20Tracing%20as%20a%20methodology%20for%20evaluating%20small%20sample%20sizes.pdf (accessed 25 November 2022).

Barrett, J. B., van Wessel, M. and Hilhorst, D. (2016) Advocacy for Development: Effectiveness, Monitoring and Evaluation.

Cartwright, N., Charlton, L., Juden, M., Munslow, T. and Williams, R.B. (2020) Making predictions of programme success more reliable. CEDIL Methods Working Paper. Oxford: Centre of Excellence for Development Impact and Learning (CEDIL).

Cartwright, N. (2020). Using middle-level theory to improve programme and evaluation design. CEDIL Methods Brief. Oxford: CEDIL.

Chambers,, R. (2007) Who counts? The quiet revolution of participation and numbers. IDS Working Paper 296. Institute of Development Studies, Brighton.

Critical Appraisal Skills Programme (CASP). (2018) CASP (insert name of checklist i.e. Qualitative) Checklist. [online] Available at: https://casp-uk.net/images/checklist/documents/CASP-Qualitative-Studies-Checklist/CASP-Qualitative-Checklist-2018_fillable_form.pdf (accessed 17 November 2022).

Copestake, J., Morsink, M. and Remnant, F. (ed.) (2019) Attributing Development Impact: the Qualitative Impact Protocol case book, Rugby, UK: Practical Action Publishing. DOI: 10.3362/9781780447469.

Ford, J.K., Schmitt, N., Schechtman, S.L., Hults, B.M. and Doherty, M.L. (1989) Process Tracing Methods: Contributions, Problems, and Neglected Research Questions. Organizational Behavior and Human Decision Processes. 43 (1), 75-117. doi:10.1016/0749-5978(89)90059-9.

Gardner, A.L. and Brindis, C.D. (2017) Advocacy and policy change evaluation: theory and practice. Stanford University Press, Stanford.

Kaiser, K., Bredenkamp, C. and Iglesias, R.M. (2016) Sin tax reform in the Philippines : transforming public finance, health, and governance for more inclusive development (English). Directions in development Washington, D.C.: World Bank Group. Available at: http://documents.worldbank.org/curated/en/638391468480878595/Sin-tax-reform-in-the-Philippines-transforming-public-finance-health-and-governance-for-more-inclusive-development (Accessed 17 September 2022).

Kamstra, J. (2017) Dialogue and Dissent Theory of Change 2.0 Supporting civil society's political role. Ministry of Foreign Affairs of the Netherlands Social Development Department Civil society unit (DSO/MO), the Hague. https://includeplatform.net/wp-content/uploads/2020/07/Annex1_DialogueandDissentTheoryofChange-June2017.pdf (accessed 20 September 2022).

Keijzer N., Spierings E., Phlix G. and Fowler A. (2011) Bringing the invisible into perspective. ECPDM. Maastricht.

Madore, A., Rosenberg, J. and Weintraub, R. (2015) "Sin Taxes" and Health Financing in the Philippines. Harvard Medical School, Brigham and Women's Hospital. http://www.globalhealthdelivery.org/case-collection/case-studies/asia-and-middle-east/sin-taxes-and-health-financing-in-the-philippines.

Mayne, J. (2020) A brief on contribution analysis: principles and concepts. 5 Oct 2020. Available from evaluatingadvocacy.org (accessed 14 October 2022).

Mayne, J. (2012) Contribution Analysis: Coming of Age? Evaluation 18 (3), 270-280.

Patton, M. (2008) Advocacy Impact Evaluation. Journal of MultiDisciplinary Evaluation, 5 (9), 1-10.

Pollard, A. and Forss, K. (2022) Evaluation Quality Assessment Frameworks: A Comparative Assessment of Their Strengths and Weaknesses. American Journal of Evaluation, 0(0). https://doi.org/10.1177/10982140211062815.

Teles, S. and Schmitt, M. (2011) 'The Elusive Craft of Evaluating Advocacy', Stanford Social Innovation Review, 9(3), pp. 38-43. DOI: 10.48558/Y90Q-VE61.

Vaessen, J., Lemire, S. and Befani, B. (2020) 'Evaluation of International Development Interventions: An Overview of Approaches and Methods', Independent Evaluation Group. Washington, DC: World Bank. Available at: http://hdl.handle.net/10986/34962.

Vigneri, M. (2021). Science in the Middle: Middle level theory in international development. CEDIL Methods Working Paper 3. London: Centre of Excellence for Development Impact and Learning (CEDIL). Available at: https://doi.org/10.51744/CMWP3

Sharma Waddington, H., Wilson, D., Aloe, A., Piggott, T., Stewart, T., Tugwell, P. and Welch, V. (forthcoming). Assessment tool for appraising risk of bias in social experiments and quasi-experiments. Mimeo.

Sidel, J.T. and Faustino, J. (2019) Thinking and Working Politically in Development: Coalitions for Change in the Philippines. Manila: The Asia Foundation. http://hdl.handle.net/11540/12348.

Van Wessel, M. (2018) 'Narrative Assessment: A new approach to evaluation of advocacy for development', Evaluation, 24(4), pp. 400-418. DOI: 10.1177/1356389018796021.

White, H. (2022) 'The unfinished evidence revolution' CEDIL Working Paper.

White, H., Saran, A., Verma, A., Oprea, E. and Babudu, P. (2021) Evidence and Gap Map of Interventions to Prevent Children Getting Involved in Violence: Technical Report on the First Edition January 2021. Youth Endowment Fund and Campbell Collaboration.

White, H., Albers, B., Gaarder, M., Kornør, H., Littell, J., Marshall, Z., Mathew, C., Pigott, T., Snilstveit, B., Waddington, H. and Welch, V. (2020) Guidance for producing a Campbell evidence and gap map. Campbell Systematic Reviews. 2020; 16:e1125. https://doi.org/10.1002/cl2.1125.

White, H. and Phillips, D. (2012) 'Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework,' 3ie Working Paper Series 15. New Delhi: International Initiative for Impact Evaluation.

Wilson-Grau, R. and Britt, H. (2013) Outcome harvesting. May 2012 (Revised November 2013). Ford Foundation, Cairo.https://www.outcomemapping.ca/download/wilsongrau_en_Outome%20Harvesting%20Brief_revised%20Nov%202013.pdf (accessed 17 July 2022).

Wilson-Grau, R. (2019) Outcome harvesting: principles, steps and evaluation approaches. Information Age Publishing.

Annexes

Annex 1: List of included programmes

Policy area	*Programme name*	*Lead organisation*	*Countries*	*Budget, (Euro millions)*	*Confidence rating*
D&D	Freedom from Fear	Pax	13	59.50	High
D&D	Fair Green and Global Alliance	Both ENDS	36	59.50	Low
D&D	Shared Resources, Joint Solutions	IUCN	16	59.50	Low
D&D	Towards a worldwide influencing network	Oxfam Novib	23	77.80	High
D&D	Building Capacity for Sector Change	UTZ	9	18.30	Medium
D&D	PRIDE	COC	16	18.30	Low
D&D	Garment Supply Chain Transformation	Fair Wear Foundation	9	32.10	Low
D&D	Health Systems Advocacy 4 Africa	Amref	5	32.10	Medium
D&D	Conducive environments for effective policy	NIMD	14	32.10	Low
D&D	No News is Bad News	Free Press Unlimited	21	32.10	Low
D&D	Advocacy for Change	Solidaridad	6	32.10	Low
D&D	GAGGA	FCAM	31	32.10	Medium
D&D	Count Me In!	Mama Cash	31	32.10	Low
D&D	Girls Advocacy Alliance	Plan Nederland	16	41.20	Low
D&D	Green Livelihoods Alliance	Milieudefensie	9	41.20	Low
D&D	PITCH	Aids Fonds	17	41.20	Medium
D&D	Partners for Resilience (PfR)	Rode Kruis	10	43.10	Low
D&D	Citizens Agency Consortium	Hivos	20	50.40	Medium/Low[7]
D&D	Prevention Up Front	GPPAC	38	10.00	Medium
D&D	Watershed - empowering citizens	IRC	6	16.40	Medium
D&D	Every Voice Counts	Care	6	16.40	Medium
D&D	SNV and IFPRI alliance	SNV	7	34.70	Medium
D&D	Empowering People in Fragile Contexts	Cordaid	9	34.70	Medium
D&D	Right Here, Right Now	Rutgers	18	34.70	Medium
D&D	Civic Engagement Alliance	ICCO	14	34.70	Medium
SRHR	Bridging the Gaps	Aids fonds	11	50.00	Medium
SRHR	Yes I Do	Plan	5	27.60	Medium
SRHR	GUSO	Rutgers	7	39.60	Low
SRHR	Jeune S3	Cordaid	4	29.80	Medium
SRHR	Her Choice	Stichting Kinderpostzegels	10	18.70	Medium
SRHR	Down to Zero	Terre des Hommes	10	15.00	Medium
SRHR	More than brides	Save the Children	10	58.60	Medium/Low[8]

[7] The reports for "Decent Work for Women" and "Open up Contracting" were rated as at 'medium confidence', while "Green and Inclusive Energy" and "Sustainable Diets for All" are rated as at 'low confidence'.

[8] The report on India, Malawi, Mali and Niger was rated as at 'medium confidence', and the Pakistan country report was rated as at 'low confidence'.

Annex 2: IOB evaluation quality criteria

Original Evaluation Quality Criteria	*Updated Evaluation Quality Criteria*
1. The problem definition concisely formulates the criteria on which the subject is to be evaluated. The evaluation questions arise from the problem definition. 2. Unambiguous description of the benchmark criteria- such as effectiveness- that are applied in the evaluation. 3. List, description and parameters of the operational population of component activities (type, target group, location, period, organisation, financial scope, etc.) to which the evaluation results relate. 4. Relevant policy-related background information and principles, and an account of the institutional setting in which the subject of the evaluation operates. 5. Description of policy theory including the assumptions about the causal and final relationships underlying the interventions evaluated and about the input/output/outcome hierarchy. 6. Degree to which the indicators defined at the various result levels can be considered specific, measurable and time-related. 7. Degree of care with which the information sources have been selected; accuracy and transparency with which data from these sources have been analysed and processed. 8. Degree to which the conclusions are actually underpinned by the evaluation results. 9. Accurate identification and justification of the methods and techniques applied in the evaluation. 10. Degree to which data have been checked, and a range of different sources/methods used for collecting data about the same characteristics and phenomena. 11. Degree to which the conclusions drawn from the sample evaluated or case studies conducted apply to the entire evaluation population. 12. Identification and explanation of any shortcomings in the evaluation and limitations on the general applicability of the findings and conclusions. 13. Degree to which the selection and content of the information sources consulted, particularly documentation and respondents, were independent of parties with an interest in the evaluation, e.g. contracting authorities, implementing agencies and beneficiaries.	U 1. A reference group oversees the evaluation U 2. Evaluators are independent U 3. Description of the context of the intervention U 4. Description of the intervention U 5. Validation of the assumptions underpinning the ToC or result chain U 6. Description of the objective of the evaluation U 7. Delimitation of the evaluation U 8. Choice of OECD-DAC evaluation criteria to be covered U 9. Clear set of evaluation questions U 10. The research design is clearly elaborated and shows how the research results will contribute to answers to the evaluation questions U 11. The methods are appropriate to evaluate effectiveness: attribution and / or contribution (if effectiveness is an evaluation criterion/question) U 12. The methods are appropriate to evaluate efficiency (if this is an evaluation criterion/question) U 13. The indicators or result areas are appropriate to capture the planned results along the different levels in the ToC U 14. Justified choice of sample, cases and information sources (e.g. choice of countries, projects, organisations and persons) U 15. The analyses are appropriate, given the chosen research design U 16. Summary of the methodology in an evaluation matrix U 17. Sufficient independent information sources U 18. Triangulation of results from different information sources U 19. Discussion of bias U 20. Systematic, complete and transparent description of the data collection and analysis U 21. Discussion of the limitations of the evaluation U 22. Conclusions answer research questions U 23. Conclusions follow logically from the research findings U 24. Validation of draft conclusions

Original Evaluation Quality Criteria	*Updated Evaluation Quality Criteria*
14. Degree to which the evaluators operated and reported independently from parties with an interest in the evaluation, e.g. contracting authorities, implementing agencies and beneficiaries. 15. Account and explanation of the progress of the evaluation, including any modifications to the original design. 16. Checks on the design and/or conduct of the evaluation by a supervisory or steering group within or outside the organisation(s). 17. Clarity of the stated aim of the evaluation (external to the evaluation itself), for which the evaluation results will be, or have been, used. 18. Degree of clarity and completeness with which the essence of the evaluation (especially its main findings) are reflected in the evaluation report and its summary. 19. Extent to which the conclusions fully answer all the evaluation questions.	U 25. Recommendations should be useful and practical, given the evaluation objectives and its intended users U 26. The report is well readable, consistent, and includes a clear summary with evaluation objective, evaluation questions, conclusions and recommendations

Annex 3: Assessment coding form

#	*Signalling question*	*Notes on signalling question*	*Responses*	*Included in score?*
#10 The research design is clearly elaborated and shows how the research results will contribute to answers to the evaluation questions				
10.1	Are the interventions of interest named or identified?		Y/PY/PN/N/UC/NA	
10.2a	Are the capacity building interventions clearly described, including implementation timelines?		Y/PY/PN/N/UC/NA	Yes
10.2b	Are the L&A interventions clearly described, including implementation timelines?		Y/PY/PN/N/UC/NA	Yes
10.2c	SRHR only: Are the service delivery interventions clearly described, including implementation timelines?		Y/PY/PN/N/UC/NA	Yes
10.3	Are the outcomes of interest clearly defined? List all outcomes with definitions.		Y/PY/PN/N/UC/NA	Yes
10.4	Is the intervention context described adequately, including contextual/external factors, such as social/cultural setting, political or economic factors, and parallel interventions or other stakeholder actions?		Y/PY/PN/N/UC/NA	
10.5	Are programme participants and project-affected persons clearly described?		Y/PY/PN/N/UC/NA	
10.6	Are the evaluation questions regarding effectiveness clearly stated?		Y/PY/PN/N/UC/NA	Yes
10.7a	What approaches do the evaluators say they planned to use to measure attribution or contribution of capacity building intervention(s) to outcome(s)?		Open-ended question	
10.7b	What approaches do the evaluators say they planned to use to measure attribution or contribution of L&A intervention(s) to outcome(s)?		Open-ended question	
10.7c	What approaches do the evaluators say they planned to use to measure attribution or contribution of service		Open-ended question	

#	*Signalling question*	*Notes on signalling question*	*Responses*	*Included in score?*
	delivery intervention(s) to outcome(s)?			
10.8	What approaches have they actually used to assess attribution?	*Using 'large n' approaches, by measuring outcomes with respect to a comparison group, using a method like difference-in-differences*	Open-ended question	
10.9	What approaches have they actually used to measure contribution?	*Using 'small n' approaches like Realist Evaluation, GEM, Process Tracing, Contribution Analysis, MSC, SCM, Outcome Mapping, MAPP, or something else?*	Open-ended question	
10.10	Does the approach belong to Group 1 (more explicit causal identification) or Group2 (more participatory approach), as defined by White & Phillips (2012)	*Group 1: realist evaluation, general elimination methodology, process tracing, contribution analysis* *Group 2: most significant change, success case method, outcome mapping, outcome harvesting, MAPP*	Y/PY/PN/N/UC/NA	
#11 The methods are appropriate to evaluate effectiveness: attribution and / or contribution				
11.1	Is the causal claim clearly stated?	*For example, "XXX caused/led to/ contributed to/impacted/affected YYY"… "Without XXX, YYY might not have happened…"/ "Otherwise, YYY would have not been possible... "*	Y/PY/PN/N/UC/NA	
11.2	Is the effect on the outcomes observed and reported?		Y/PY/PN/N/UC/NA	
11.3	Is a change in outcomes observed relative to a comparison group (that is, a group that does not receive the intervention of interest)?		Y/PY/PN/N/UC/NA	
11.4	Is there a timeline showing that the cause (implementation of the intervention) preceded the event (observed change in outcome)?		Y/PY/PN/N/UC/NA	Yes
11.5	Is there a plausible posited causal mechanism underlying the relationship between intervention and outcome?		Y/PY/PN/N/UC/NA	Yes

#	*Signalling question*	*Notes on signalling question*	*Responses*	*Included in score?*
11.6	Does the evaluation articulate alternative causal hypotheses, including the role of contextual/external factors, such as social/cultural setting, political or economic trends, and parallel interventions or other stakeholder actions, that may influence outcomes?		Y/PY/PN/N/UC/NA	Yes
11.7	Is the qualitative methodology, which will interrogate the relationship between intervention and outcome, described?		Y/PY/PN/N/UC/NA	Yes
#13 The indicators or result areas are appropriate to capture the planned results along the different levels in the ToC				
13.1	Is the ToC presented for the intervention(s) being evaluated, that - sets out underlying intervention logic and theoretical links - outlines inputs, activities, outputs, intermediate and final intended outcomes - lists programme participants and project-affected persons, timelines and indicators to monitor change - provides assumptions and risks at each link in the chain - provides contextual factors and external influences in causal chain?	*Y if all the conditions are met; Probably Y if 3-4 met; Probably No if 1-2 met; No if none of them are met.*	Y/PY/PN/N/UC/NA	Yes
13.2	Are measurable indicators presented for the intervention(s) being evaluated (Impact-Outcome-Output-Activities-Inputs), for example in a log frame?		Y/PY/PN/N/UC/NA	Yes
13.4	Does the ToC/log frame articulate possible unintended outcomes (e.g., spillovers)?		Y/PY/PN/N/UC/NA	Yes
13.5	Is the selection of outcome collected/changes observed justified with reference to the ToC or otherwise?		Y/PY/PN/N/UC/NA	Yes
13.6	What measures or measurement instruments or		Open-ended question	

#	*Signalling question*	*Notes on signalling question*	*Responses*	*Included in score?*
	approaches are used to measure capacity of CSOs?			
#14 The choice of sample, cases and information sources is justified				
14.1	Is a stakeholder map presented ('omitted informant bias')?		Y/PY/PN/N/UC/NA	Yes
14.2	Is the list of interviewees presented (e.g., in an appendix)?		Y/PY/PN/N/UC/NA	
14.3	Is the list of documents presented (e.g., in an appendix)?		Y/PY/PN/N/UC/NA	
14.4	Does the recruitment or sampling strategy describe how have the participants been selected?		Y/PY/PN/N/UC/NA	
14.5	Is the sample selection process explained and justified?		Y/PY/PN/N/UC/NA	Yes
14.6	Are sample characteristics adequately reported (sample size, location, and at least one additional characteristic)?		Y/PY/PN/N/UC/NA	Yes
14.7	Is the recruitment or sampling strategy appropriate, including explaining why the participants selected were the most appropriate to provide access to the knowledge sought to answer the evaluation questions?		Y/PY/PN/N/UC/NA	
#15 The analyses are appropriate, given the chosen research design				
15.1	Is there a detailed description of the analysis process?		Y/PY/PN/N/UC/NA	Yes
15.2	Contribution analysis: according to what is reported, is the method implemented appropriately?	*Contribution analysis involves: 1) articulating ToC; 2) evaluating whether intervention activities implemented as set out; 3) chain of expected results (outcomes) shown as having occurred; 4) other influencing factors ruled out or relative contribution recognised.*	Y/PY/PN/N/UC/NA	
15.3	Outcome mapping: according to what is reported, is the method implemented appropriately?	*Outcome mapping involves: 1) articulating ToC "intentional design" and "boundary partners"; 2) collection of outcome, strategy and*	Y/PY/PN/N/UC/NA	

#	*Signalling question*	*Notes on signalling question*	*Responses*	*Included in score?*
		performance journals, which may incorporate Most Significant Change (MSC) analysis; 3) "evaluation planning" (data collection and verification)		
15.4	Most significant change: according to what is reported, is the method implemented appropriately?	*Most Significant Change (MSC) involves: 1) defining domains of change and timeframe; 2) systematic collection of stories from participants about (positive and negative) changes that occurred in their lives in the recent past, enquiries about why the changes occurred and were significant; 3) systematic review of stories of change by stakeholder panels; 4) verification of stories through additional data collection and possible quantification of changes; 5) comparison of most significant change stories with expected changes in ToC/log-frame*	Y/PY/PN/N/UC/NA	
15.5	Outcome harvesting: according to what is reported, is the method implemented appropriately?	*Outcome harvesting involves: 1) gathering data on potential outcomes to which change agent may affect and contributions by change agent; 2) verification through informant review of draft outcomes, usually in workshop, and evaluator assessment of plausibility and coherence; 3) substantiation of outcomes and contributions through additional data interviews; 4) categorisation and interpretation of outcomes*	Y/PY/PN/N/UC/NA	
15.6	Is the data analysis approach presented in sufficient detail and justified?		Y/PY/PN/N/UC/NA	Yes
#16 Summary of the methodology in an evaluation matrix				
16	Does the study present an evaluation matrix or plan linking evaluation questions with nature and sources of data, protocols for qualitative		Y/PY/PN/N/UC/NA	Yes

#	*Signalling question*	*Notes on signalling question*	*Responses*	*Included in score?*
	field work and categories for data analysis?			
#17 Sufficient independent information sources				
17.1	Are separate types of information sources used e.g., documents, interviews, focus groups, field visits?		Y/PY/PN/N/UC/NA	Yes
17.2	What are these separate sources of information (Government statistics, surveys conducted my other entities, etc.)?		Open-ended question	
17.3	Does the data collection attempt to guard against cherry picking of cases, such as through random sampling of targeted programme participants or purposive sampling across a diverse group using a sampling frame (e.g., including those who may have dropped out), or indicate methods taken to avoid convenience sampling of respondents?		Y/PY/PN/N/UC/NA	
17.4	Are appropriate sources included that were involved in delivering or receiving the intervention - e.g., participants, implementers, programme managers?		Y/PY/PN/N/UC/NA	Yes
17.5	Are relevant sources included that were not involved in, or may have experienced another, intervention - e.g., trade union members?		Y/PY/PN/N/UC/NA	Yes
17.6	Is there discussion of issues around recruitment (e.g., why some people chose not to take part)?		Y/PY/PN/N/UC/NA	Yes
#18 Triangulation of results from different information sources				
18.1	Is the evidence of a causal relationship triangulated?		Y/PY/PN/N/UC/NA	Yes
18.2	Describe the method(s) of triangulation used	*-Data triangulation (location, time and participants)* *-Investigator triangulation* *-Theory triangulation (several theories)* *-Methodological triangulation*	Open-ended question	

#	*Signalling question*	*Notes on signalling question*	*Responses*	*Included in score?*
18.3	Are these methods appropriate to answer evaluation questions?		Y/PY/PN/N/UC/NA	
#19 Discussion of bias				
19.1	Are possible alternative causal chains/claims presented?		Y/PY/PN/N/UC/NA	Yes
19.2	Does the study attempt to rule out alternative explanations for changes in outcomes, such as analysis of alternative hypotheses or falsification methods (irrelevant interventions or outcomes)?		Y/PY/PN/N/UC/NA	Yes
19.3	Is the evaluator's own position, assumptions and possible biases discussed, in order to protect against evaluator bias (e.g., 'friendship'/'contract renewal bias')?		Y/PY/PN/N/UC/NA	Yes
19.4	Is the evaluator affiliation financially independent from the organization being evaluated?		Y/PY/PN/N/UC/NA	Yes
19.5	Does the study attempt to protect against respondent bias*?	*e.g.,* *- by drawing up questions to avoid leading questions in interviews* *- BLINDING participants to the evaluation* ** respondent bias includes: includes courtesy bias/ political correctness bias, positional bias (e.g. errors of attribution to intervention), self-serving bias, self-importance bias*	Y/PY/PN/N/UC/NA	Yes
19.6	Are the data collected within a sufficiently short time period from implementation of the intervention to protect against recall bias (e.g., interviews conducted while the programme is ongoing)?		Y/PY/PN/N/UC/NA	Yes
19.7	Does the study attempt to protect against evaluator bias by recording interviews and comparison of notes by multiple interviewers (confirmation bias)?		Y/PY/PN/N/UC/NA	Yes

#	*Signalling question*	*Notes on signalling question*	*Responses*	*Included in score?*
19.8	Was the potential for conflict of interest considered and addressed?		Y/PY/PN/N/UC/NA	Yes
#20 Systematic, complete and transparent description of the data collection and analysis				
20.1	For factual information: are initial themes, categories and data codes structured around ToC/log-frame/results framework?		Y/PY/PN/N/UC/NA	
20.2	For counterfactual information: are data collection protocols linked to comparison groups or possible alternative hypotheses?		Y/PY/PN/N/UC/NA	
20.3	Is it clear how the data were collected from informants; e.g. is there a discussion of how interviews/FGDs were conducted and recorded?		Y/PY/PN/N/UC/NA	Yes
20.4	Is it clear how document reviews were conducted; e.g. is a data collection sheet containing codes presented?		Y/PY/PN/N/UC/NA	Yes
#21 Discussion of the limitations of the evaluation				
21.1	Do the findings address the evaluation questions?		Y/PY/PN/N/UC/NA	
21.2	Are all potential limitations thoroughly discussed (limitation due to data availability, resource (time/funds) constraints, risk of bias and any other sorts of limitation)?		Y/PY/PN/N/UC/NA	Yes
21.3	Are the implications or recommendations clearly linked to/based on the evidence from the study?		Y/PY/PN/N/UC/NA	
21.4	Does the research comply with ethics: anonymity, informed consent, and confidentiality.		Y/PY/PN/N/UC/NA	Yes

Outcome data collection protocol

Evaluation #	Theme (D&D only)	Sub-group (e.g., country)	Measured change (+, 0, -)	Contribution (reported) (Strong/Medium/ Weak/Unclear)	Evidence rating (reported) (Strong/Medium/ Weak/Unclear)	Description (provide page numbers and text)	Outcome category

Note: outcome categories

(1) D&D programmes

- Capacity development activities (or outputs achieved) with partner CSOs
- Capacity development activities (or outputs achieved) with other stakeholders
- Support to L&A activities (or outputs achieved) by partner CSOs
- L&A activities (or outputs achieved) by other stakeholders
- Skills/capacities of local partners/CSOs
- Spillovers to skills/capacities to other local CSOs
- Partnership/coalition building/collaborations with other actors
- L&A activities by local partners/CSOs
- Community-level outcomes
- Policy engagement
- Policy change outcomes
- Policy implementation outcomes

(2) SRHR Programmes

- Activities completed/outputs achieved
- Knowledge/ information
- Girls' attitudes
- Attitudes of other community members
- Girls' empowerment (e.g., involvement in decision making)
- Access to SRHR services
- Access to complementary services
- SRHR service use
- Sexual and reproductive health outcomes

Annex 4: Outcomes harvested

Measurement of capacity development

Outcome sub-category	*Examples of outcome measures*	*Data collection method*
Capacity development activities (or outputs achieved) with partner CSOs	Description of outputs (e.g., CSOs participated in the workshops provided by the programme)	Desk review
	Number or % of partners that received capacity development support	Survey
Capacity development activities (or outputs achieved) with other stakeholders	Training completed (e.g., "capacity building of MPs on roles and missions and the use of ICT to interact with citizens was conducted")	Field study, interview and/or survey
	Number or % of stakeholders who feel their capacity was increased	Stories reported by stakeholders
	Number of stakeholders (e.g., youths) who received training	CATool, survey
Measurement of skills and capacities of local partners and CSOs	'5C Framework' (knowledge on the topic, skills to engage with private sector, capacity for evidence-based lobby, capacity for networking, skills to conduct research, outcome harvesting or monitoring)	Field visit, e-survey
	Capacity Self-Assessment	Document review, interviews

Measurement of support to lobby and advocacy

Outcome sub-category	*Examples of outcome measures*	*Data collection method*
Support to L&A activities (or outputs achieved) by partner CSOs	Specific activities (e.g., community advocates launched a task force to pool together resources of the members to coordinate community-based advocacy awareness activities and to raise funds)	Analysis of harvested outcomes
	Number of advocacy initiatives supported by the programme	Document review, FGD, interview, survey
L&A activities by local partners/CSOs	Specific activities (e.g., group formed through the programme participating at advocacy events, publication of reports and research, participation public consultations)	Desk review, field interview, FGD, analysis of harvested outcomes
	Number of L&A activities carried out	Survey
	Number of evidence products in support of L&A	Document review, survey, interviews, storytelling
	Number of CSOs that started undertaking dialogues with decision makers	Document review, survey, interviews, storytelling
L&A activities (or outputs achieved) by other stakeholders	Specific activities (e.g., a working group continues to push for transitional justice, raising awareness with government and civil society)	Field study, interviews, survey and/or analysis of harvested outcomes

Measurement of skills and capacities of local partners and CSOs

Examples of outcome measures	*Data collection method*
'5C Framework' (knowledge on the topic, skills to engage with private sector, capacity for evidence-based lobby, capacity for networking, skills to conduct research, outcome harvesting or monitoring)	Field visit, e-survey
Capacity Self-Assessment	Document review, interviews
V4CP Capacity Assessment scores	Review of annual reports/ country reports, survey and interviews
Participatory Capacity (self-) Assessment Tool (PCAT)	PCAT
Stories of self-reported capacity or achievements (e.g., respondent stating that they have more competencies to conduct advocacy efforts with political decisionmakers)	Stories of capacity change collected online through Sprockler
Implementation of new approach to monitoring, documenting and reporting, and data management	Field study, interview and/or FGD, analysis of harvested outcomes
Knowledge on fiscal and budgetary matters	CA Tool
% partners who have increased capacity to generate or use verified evidence	Interviews, workshop
Internalisation or use of training content	Interviews, workshop
Number or % of partners who feel their capacity or knowledge increased	Survey, desk research, interviews
Number of respondents indicating organisation strengthened	Stories of capacity change collected online (Sprockler)

Measurement of partner capacity

Outcome sub-category	*Examples of outcome measures*	*Data collection method*
Partnerships, coalition building, collaboration with other actors	Specific stories of partnership or collaborations formed	Field study, interviews and/or FGD, analysis of harvested outcomes
	Perceived level of interaction or trust among stakeholders	Interview/ workshop
	Number of partnerships formed or strengthened	Field study, interviews and/or FGD, analysis of harvested outcomes
Spillovers to skills and capacities to other local CSOs	Specific stories about spillovers (e.g., capacity building benefited partners of partners, and outside actors such as communities and government agencies)	Analysis of harvested outcomes, survey, annual report reviews, stories of change

Measurement of knowledge, attitudes and empowerment

Outcome sub-category	*Examples of outcome measures*	*Data collection method*
Knowledge and information	Number or % of youth reported having increased access to SRHR information sources	Desk review, interviews, survey, FGD

Outcome sub-category	*Examples of outcome measures*	*Data collection method*
	% girls who know about protective laws on child marriage and FGM	Questionnaire, workshop, interview
	% children who are aware of sexual exploitation of children and its risks	Workshops and programme monitoring data
Girls' attitudes	Self-report (e.g., girls now decide themselves about the use of preventive methods or that they felt more confident saying "no" in the context of proposed sexual intercourse)	FGD, interview
	Number or % of girls agreeing that girls have a right to refuse an arranged marriage and that girls have a right to divorce	Questionnaire, workshop, interviews
Attitudes of other community members	Specific stories (e.g., a negative backlash by the media and increased instances of cyber-bullying towards LGBT following the submission of petition, youth reporting an increase in awareness with regards to SRHR)	Document review, interviews, analysis of outcome harvests
	Number of community members who see child marriage and early pregnancy as a good practice or responsibility of girls	Interviews
Empowerment	Capacity to advocate for themselves, including at community gatherings	Desk review, interviews, survey, FGD
	Girls' decision-making power and participation	Desk review, interviews, survey, FGD
	School attendance and enrolment	Questionnaire, workshop, interviews

Measurement of policy engagement and policy change

Outcome sub-category	*Examples of outcome measures*	*Data collection method*
Policy engagement	Specific engagement stories (e.g., local authorities refer to the report published through the programme, voices of female farmers better included in the development of policies)	Field study, interviews and/or FGD
	Number of strategy decision made to guide the behaviour of governmental actors	Analysis of harvested outcomes
	Number of countries that created space for CSOs and take the CSOs' cause into account in their policy and development plan	Analysis of harvested outcomes
Policy change	Specific examples of policy change (or lack of change)	Desk review, field study, interviews, FGD, analysis of harvested outcomes
	Number of policies and laws that have been adopted, blocked, maintained or introduced on the issue	Survey
	Number of countries that took steps for policy change	Analysis of harvested outcomes
Policy implementation	Specific stories of better policy implementation	Stories reported by stakeholders

Outcome sub-category	*Examples of outcome measures*	*Data collection method*
Policy engagement	Specific engagement stories (e.g., local authorities refer to the report published through the programme, voices of female farmers better included in the development of policies)	Field study, interviews and/or FGD
	Number of strategy decision made to guide the behaviour of governmental actors	Analysis of harvested outcomes
	Number of countries that created space for CSOs and take the CSOs' cause into account in their policy and development plan	Analysis of harvested outcomes
	Number of policies and laws that were adequately implemented	Document review, survey, interviews, storytelling
	Number of negative outcomes related to poor implementation of bills	Analysis of harvested outcomes and stories of change
	Number of countries that saw downstream effects as a result of better policy implementation (e.g., small scale farmers gained access to market as a result of issue of certificates)	Desk study, interviews and workshop
	Number of countries that saw better policy implementation	Analysis of outcome harvest data

Measurement of service access and use

Outcome sub-category	*Examples of outcome measures*	*Data collection method*
Access to SRHR services	Specific stories (e.g., local NGOs providing access to SRHR for the first time, sex workers have a place to go to for support)	Document review, interviews, analysis of outcome harvests
	Number of community-led clinics providing HIV prevention, treatment and care	Document review, interviews, analysis of outcome harvests
	Number of people living with HIV accessing health services in the area	Document review, interviews, analysis of outcome harvests
	Supply and availability of drugs	Document review, interviews, analysis of outcome harvests
	% children accessing SRHR services	Workshops and programme monitoring data
	% girls reporting their community has a youth-friendly health centre	Survey, interviews
Access to complementary services	Number of community health centre where staff was trained	Document review, interviews, analysis of outcome harvests
	Number of public health centres that have become referral centre	Document review, interviews, analysis of outcome harvests

Outcome sub-category	*Examples of outcome measures*	*Data collection method*
	Specific stories (e.g., mental health service has been integrated into public health care, referral system is functioning)	Survey, interviews
Service use	% health staff confirming youth increased their attendance at health centres	Interviews with health staff
	Number of respondents indicating they had ever used SRHR services	Survey, interviews
	Number of young people visiting SRHR services	Interviews, FGD, workshop, field visits

Measurement of community-level outcomes

Examples of outcome measures	*Data collection method*
Thematic outcomes (e.g., reduction in violence and conflict in the community) Specific events (e.g., ceremony of conflict resolution held)	Field study, interviews and/or FGD
Community awareness, stories of norms changed (e.g., description of what happened, perceived level of knowledge)	Field study, interviews and/or FGD
Media attention (e.g., public expressions in favour of SRHR)	Stories reported by stakeholders
Religious and traditional leaders' commitments (e.g., declaration at a specific event)	Field study, interviews, survey and/or analysis of harvested outcomes
Description of empowerment output (e.g., self-help group formed) in target population (e.g., youth, workers)	Field study, interviews, survey and/or analysis of harvested outcomes
CSO engagement description (e.g., creation of neutral space for CSOs to have conversations with policy makers)	Document review, interviews

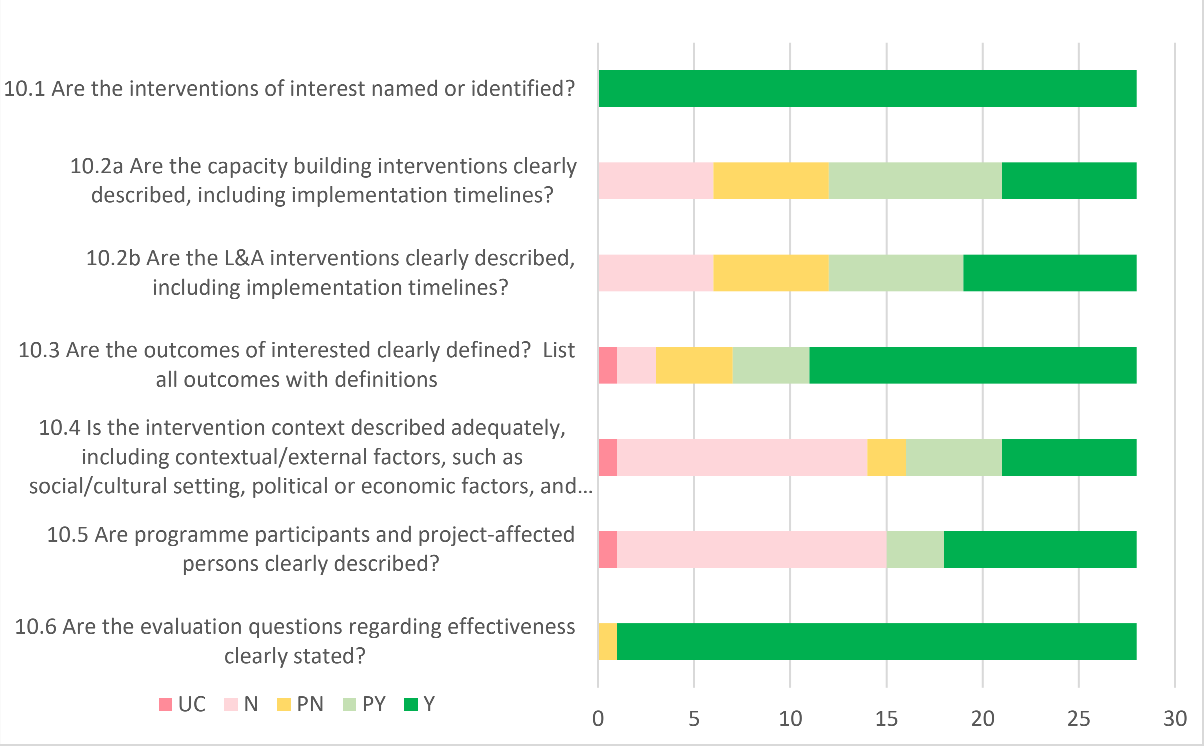
Measurement of sexual and reproductive health outcomes

Examples of outcome measures	*Data collection method*
Number of STI cases, new HIV infections	Document review, interviews, analysis of outcome harvests
HIV treatment coverage	Document review, interviews, analysis of outcomes harvested
Number of cases of child marriage	Desk review, interviews, survey, FGD
Reported practice of FGM/C	Desk review, interviews, survey, FGD
Frequency of sexual harassment in school	Desk review, interviews, survey, FGD
% contraceptive use among young people	Survey, interviews

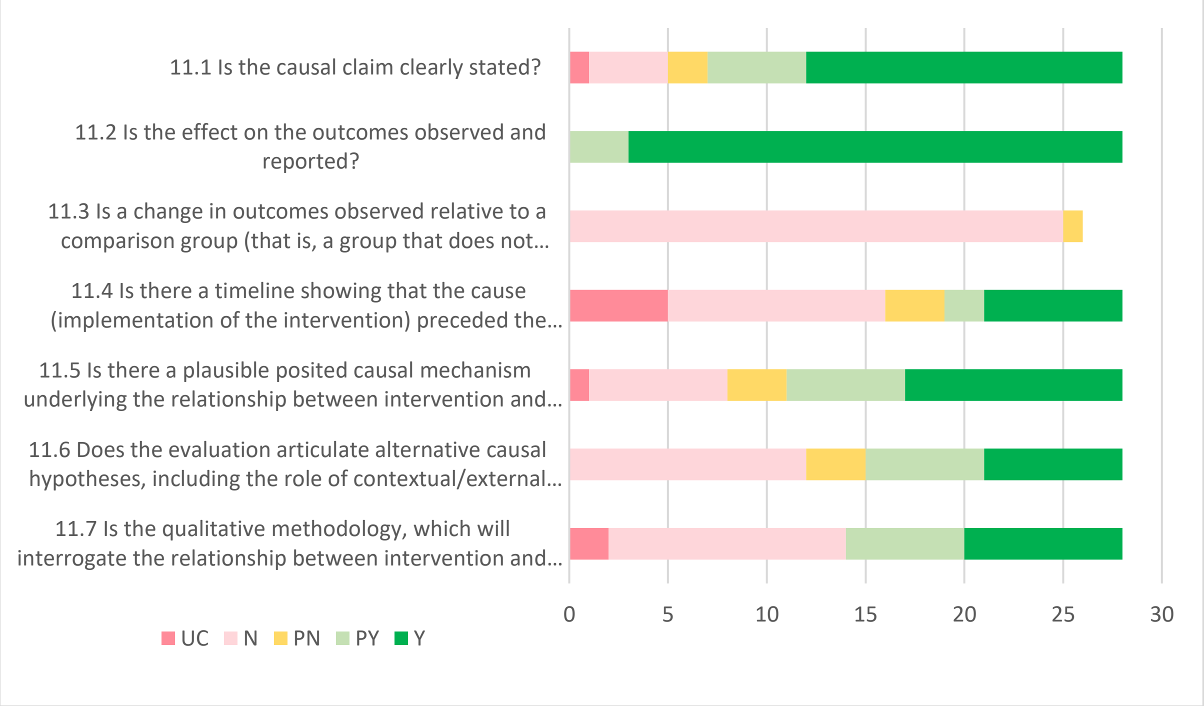
Annex 5: Results of assessments

Assessment findings for D&D

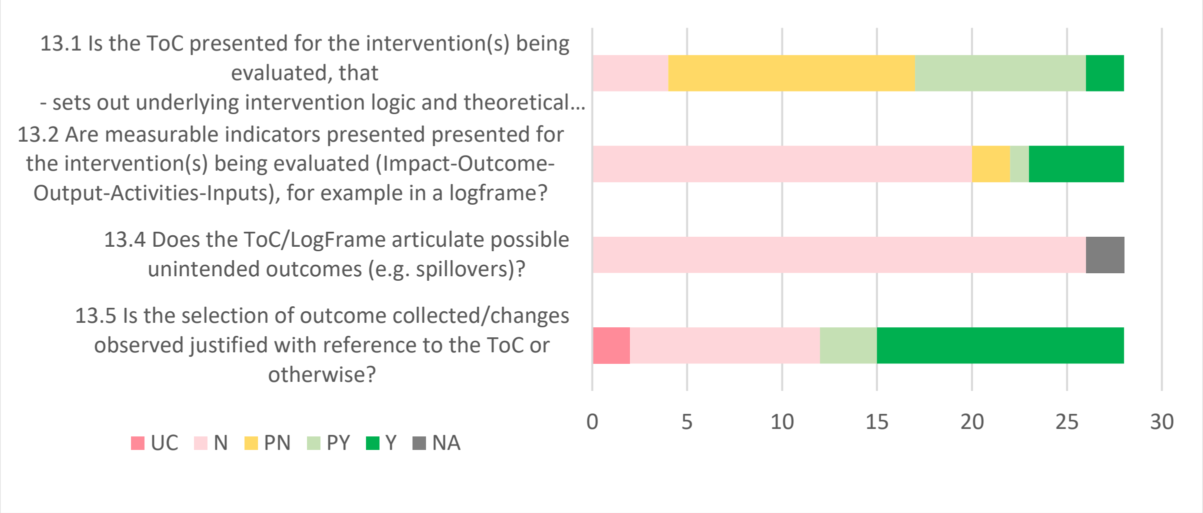
Criterion 10 "Research design"



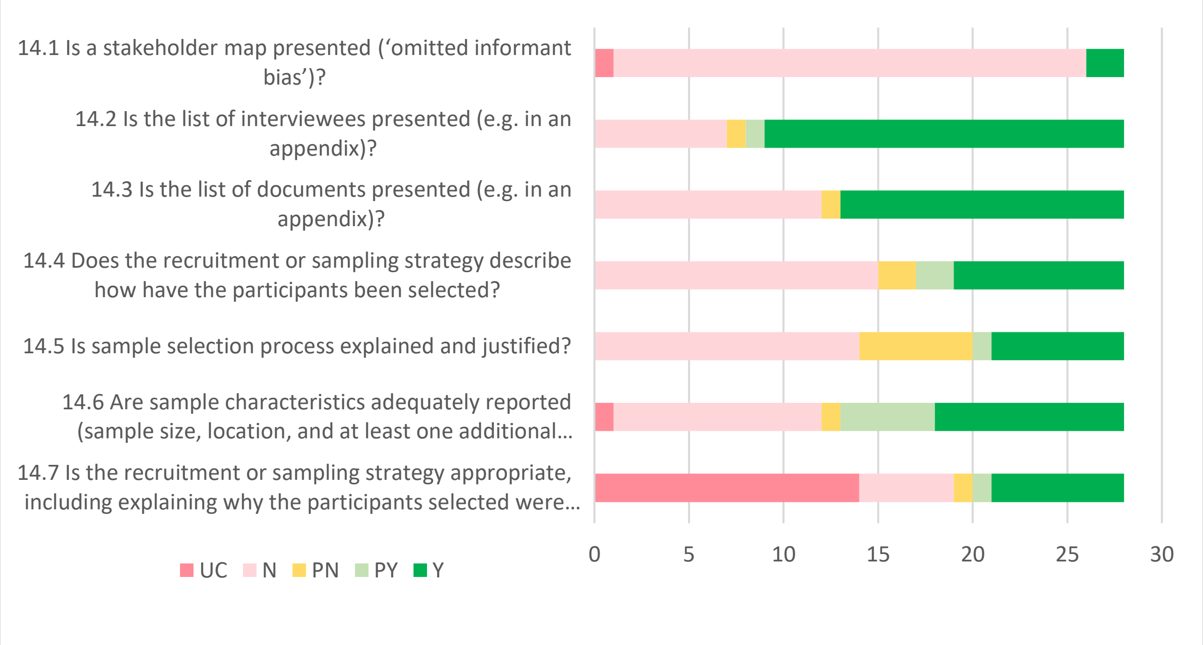
Criterion 11 "The methods are appropriate to evaluate effectiveness"



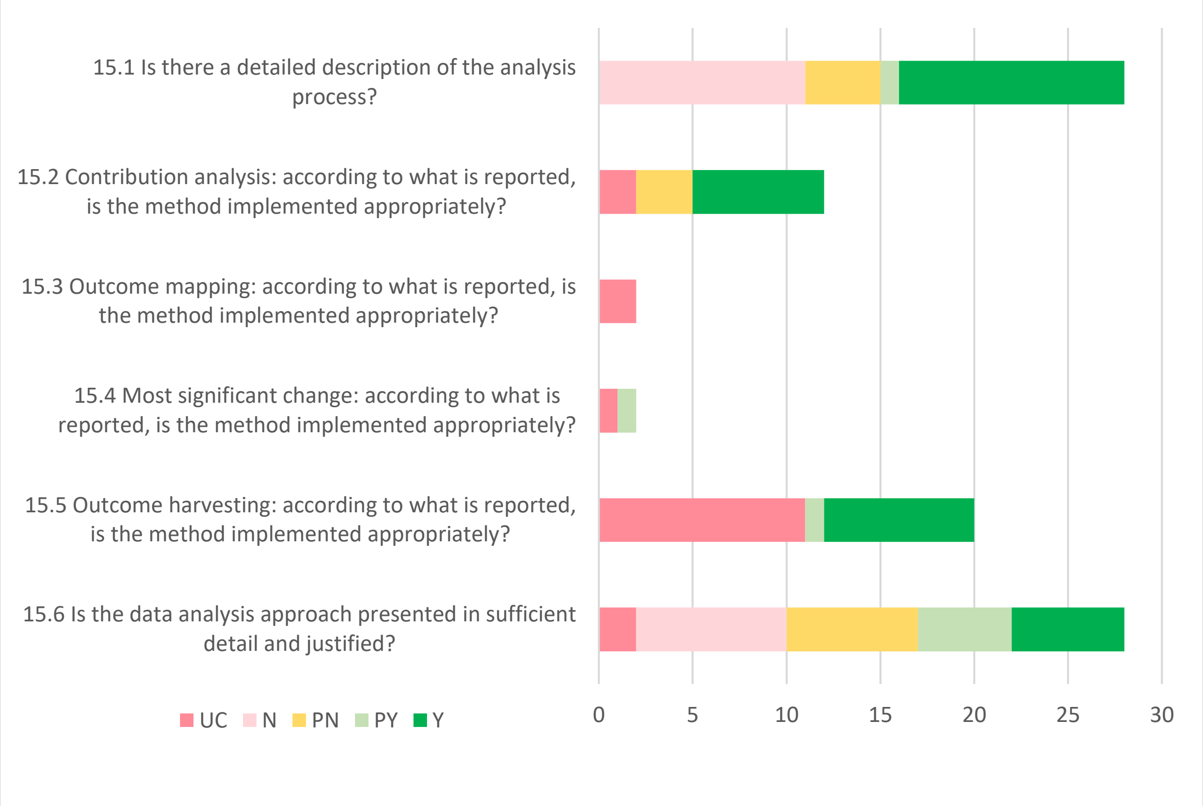
Criterion 13 "The indicators or result areas are appropriate to the ToC"



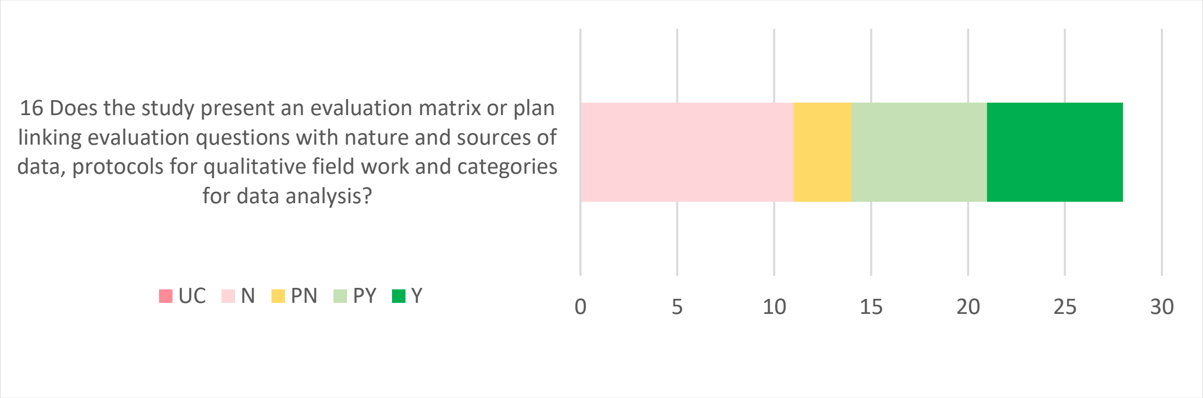
Criterion 14 "Justified choice of sample, cases and information sources"



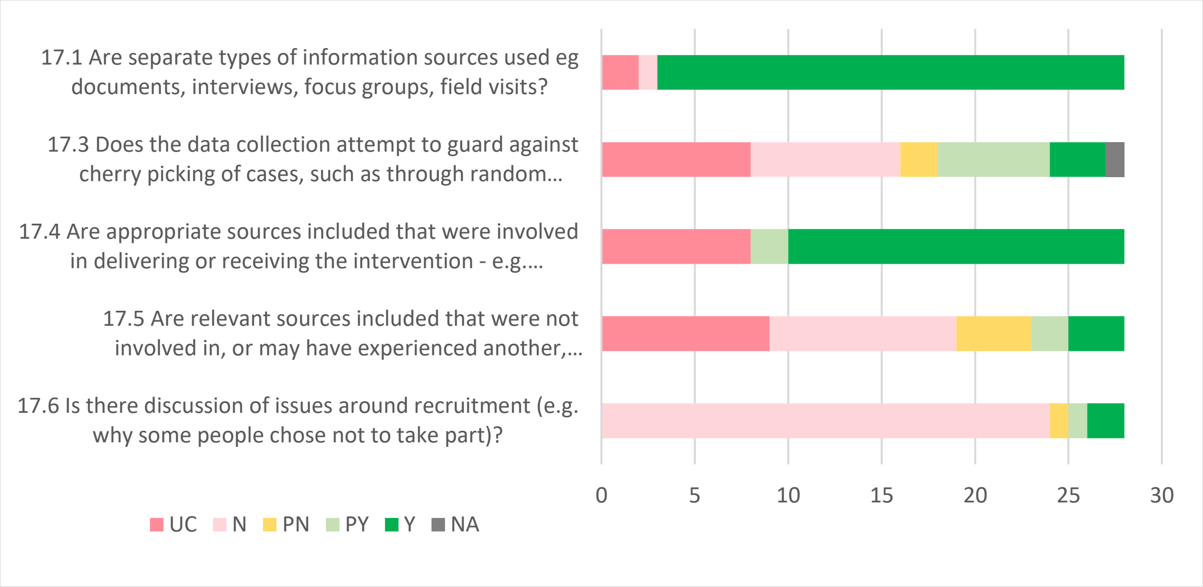
Criterion 15 “The analyses are appropriate, given the chosen research design”



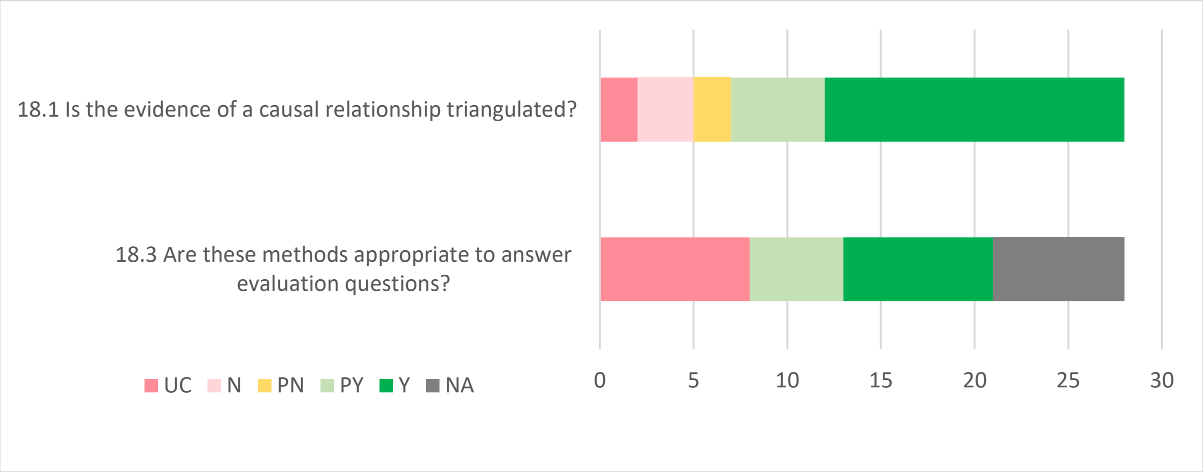
Criterion 16 “Summary of the methodology in an evaluation matrix”



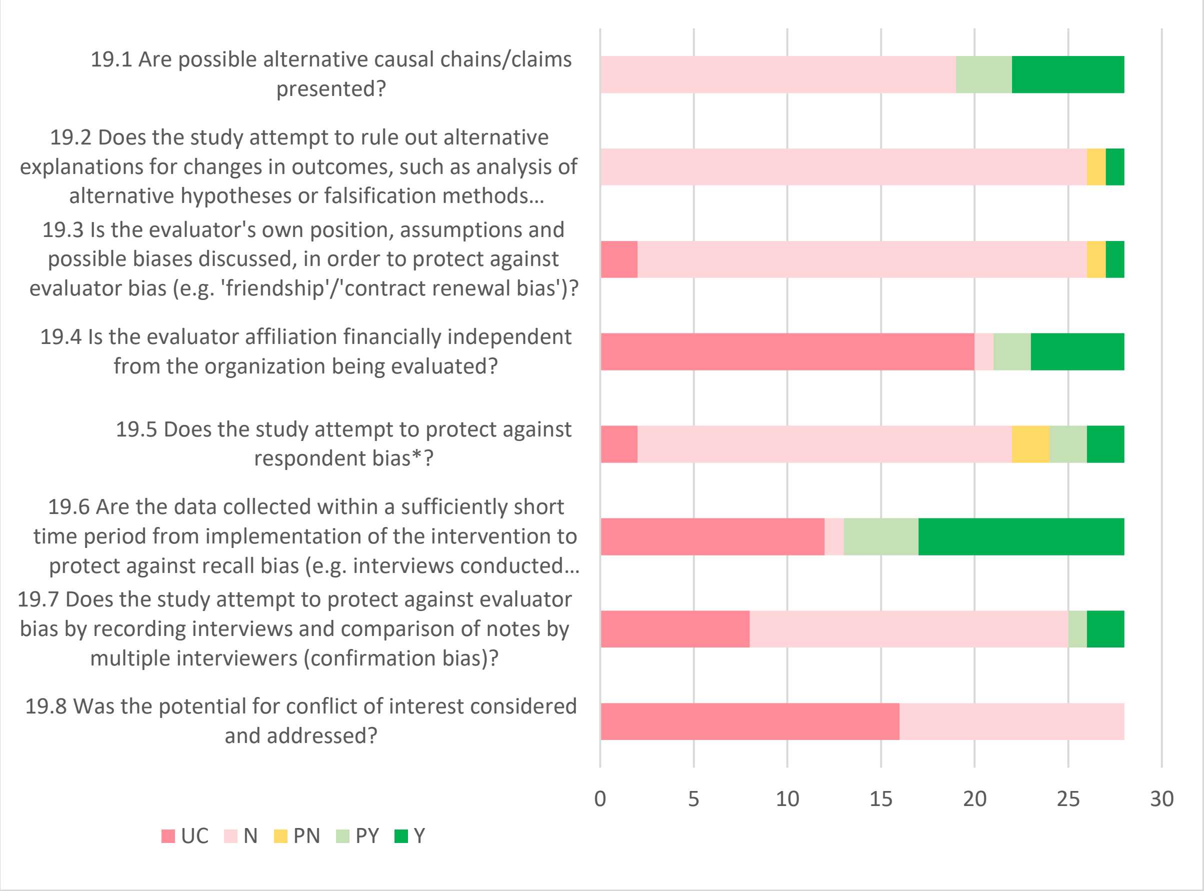
Criterion 17 “Sufficient independent information sources”



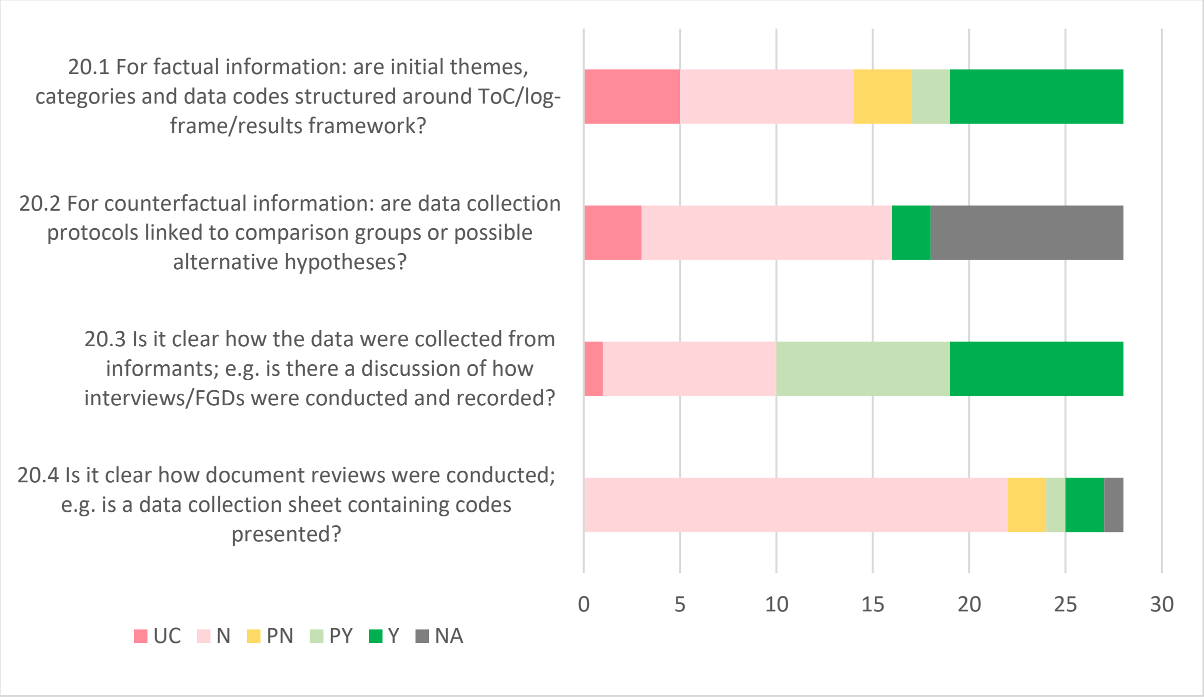
Criterion 18 “Triangulation of results from different information sources”



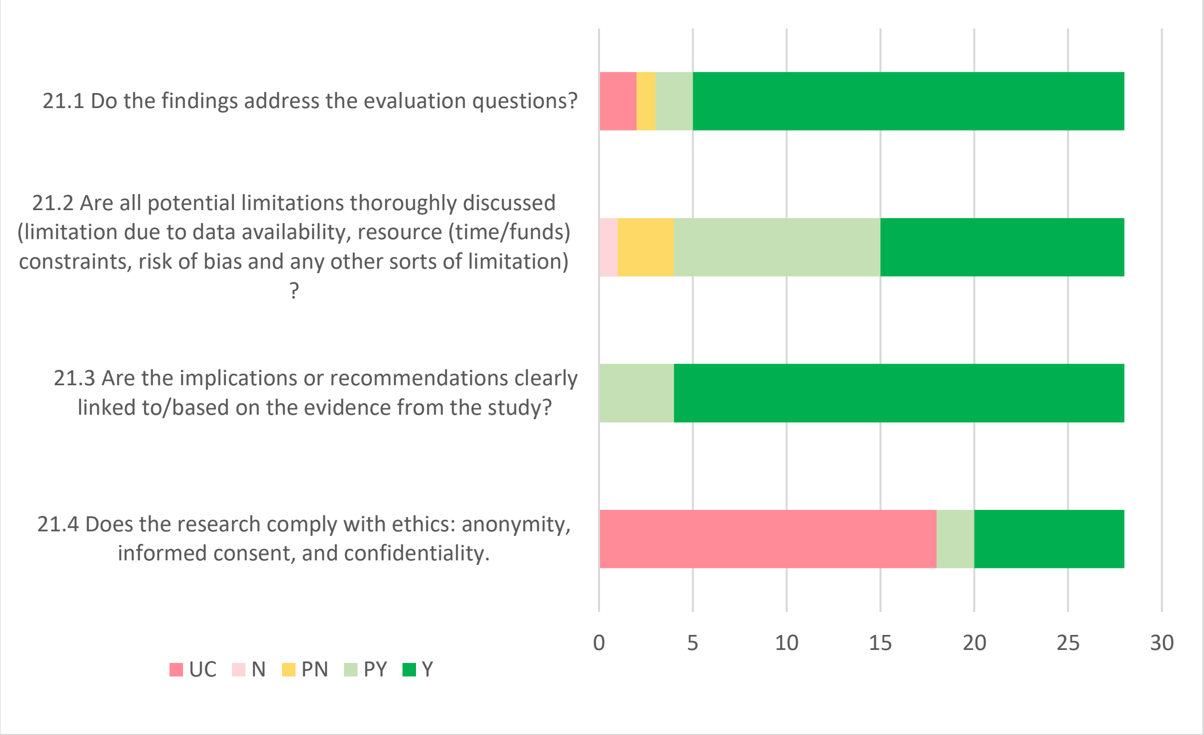
Criterion 19 “Discussion of bias”



Criterion 20 “Systematic, complete and transparent description of the data collection and analysis”

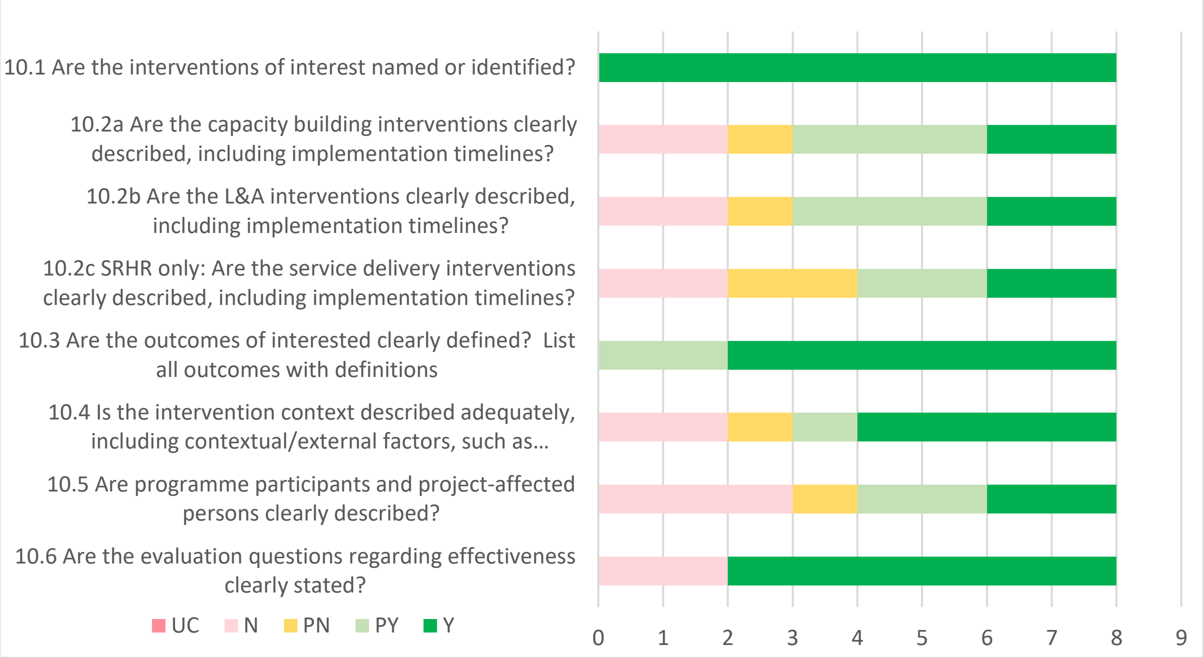


Criterion 21 “Discussion of the limitations of the evaluation”

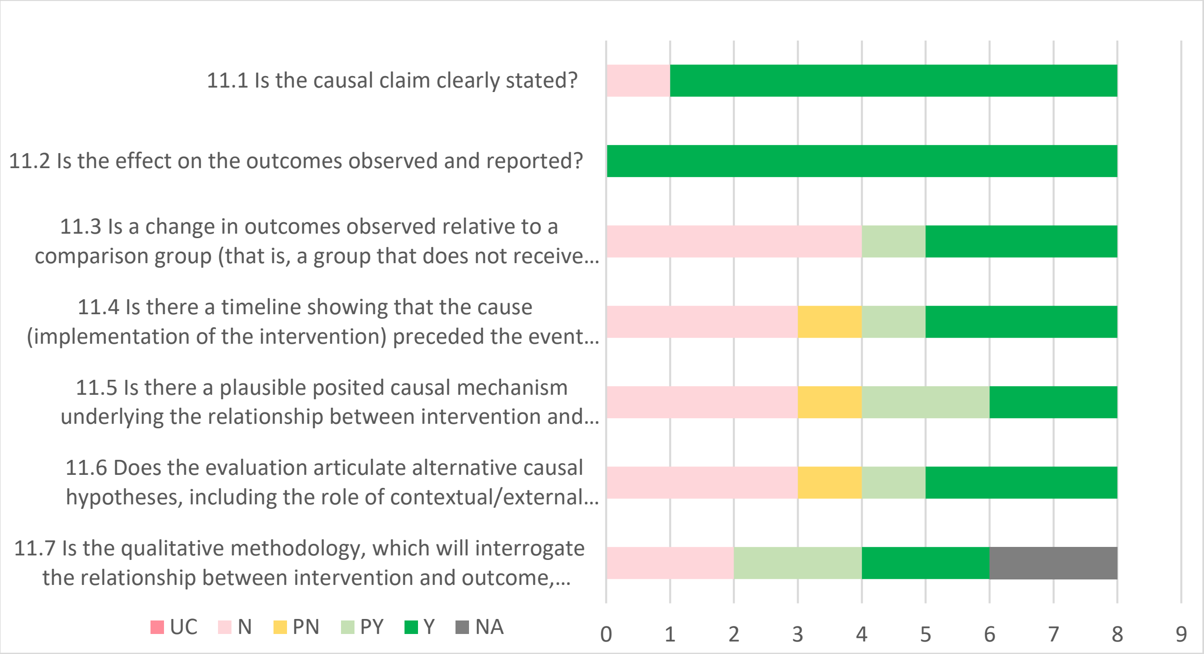


Assessment findings for SRHR

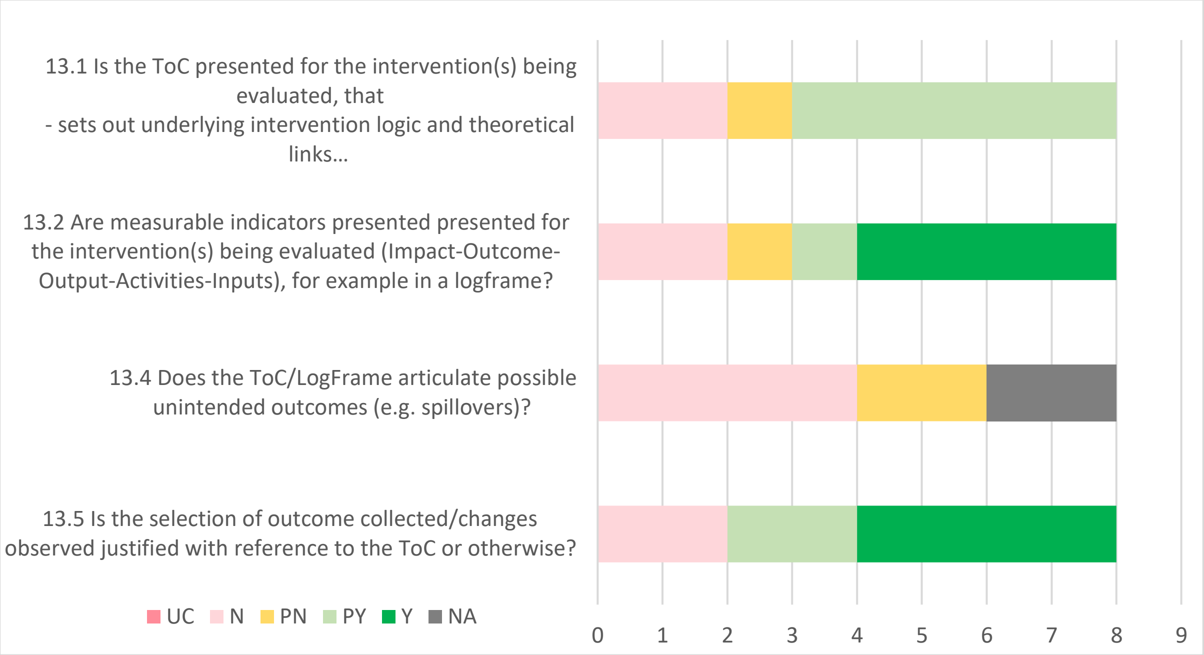
Criterion 10 "Research design"



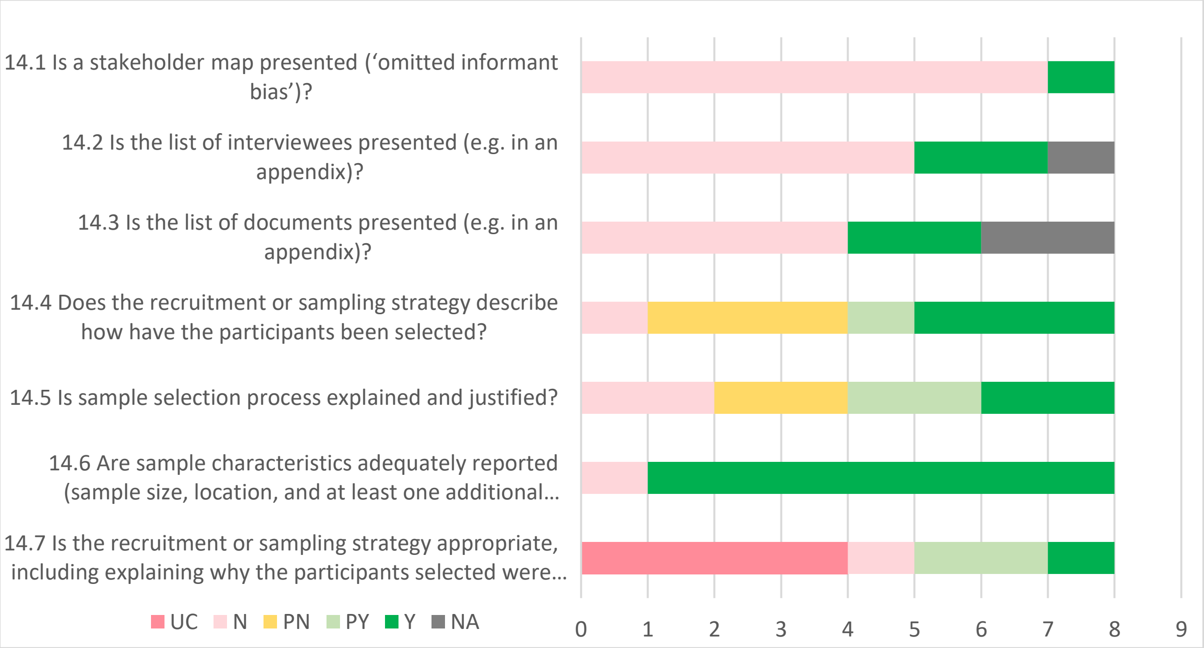
Criterion 11 "The methods are appropriate to evaluate effectiveness"



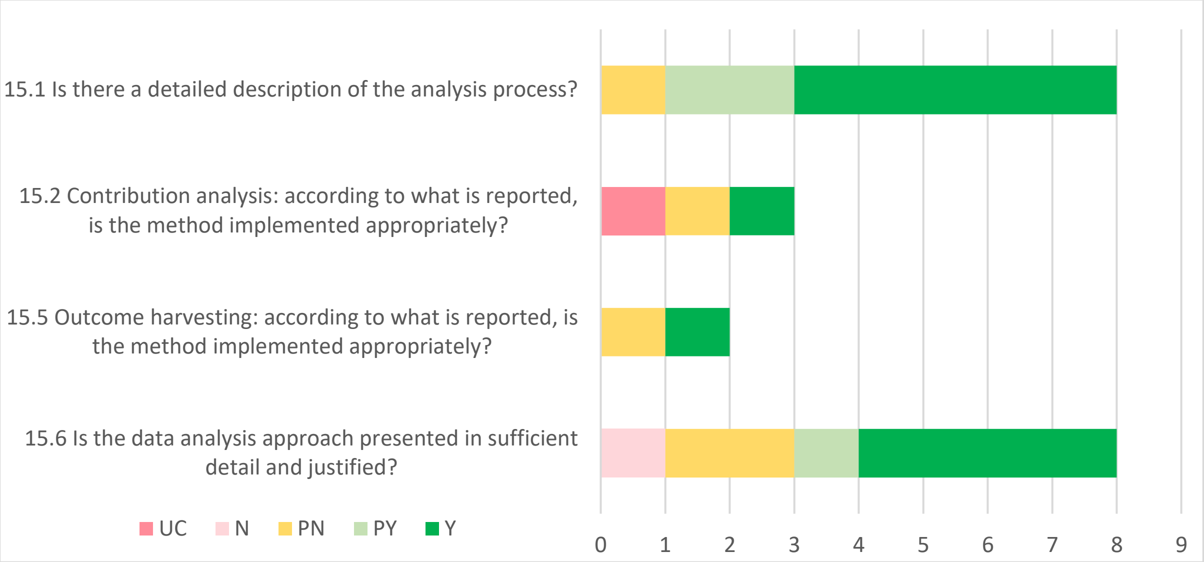
Criterion 13 “The indicators or result areas are appropriate to the ToC”



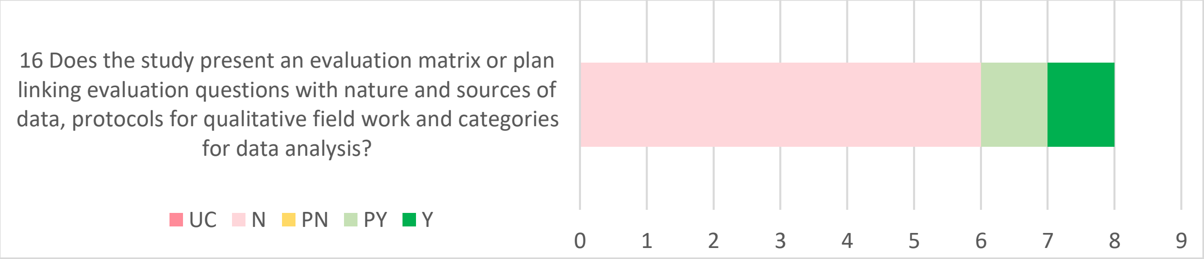
Criterion 14 “Justified choice of sample, cases and information sources”



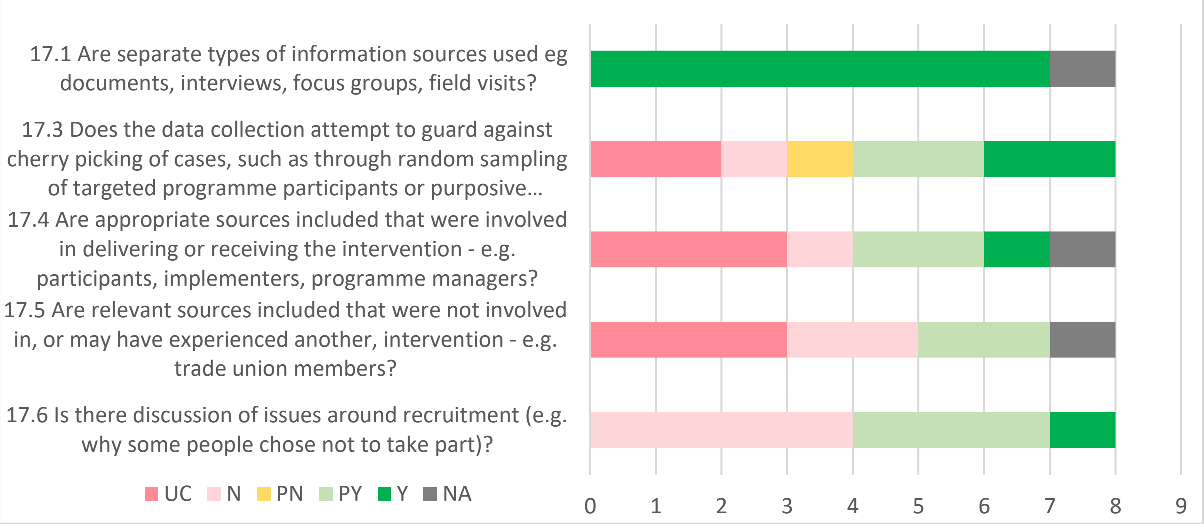
Criterion 15 “The analyses are appropriate, given the chosen research design”



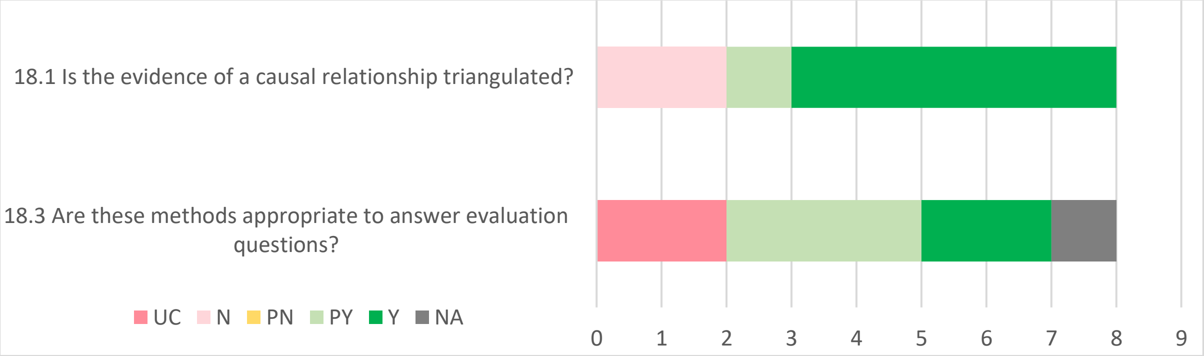
Criterion 16 “Summary of the methodology in an evaluation matrix”



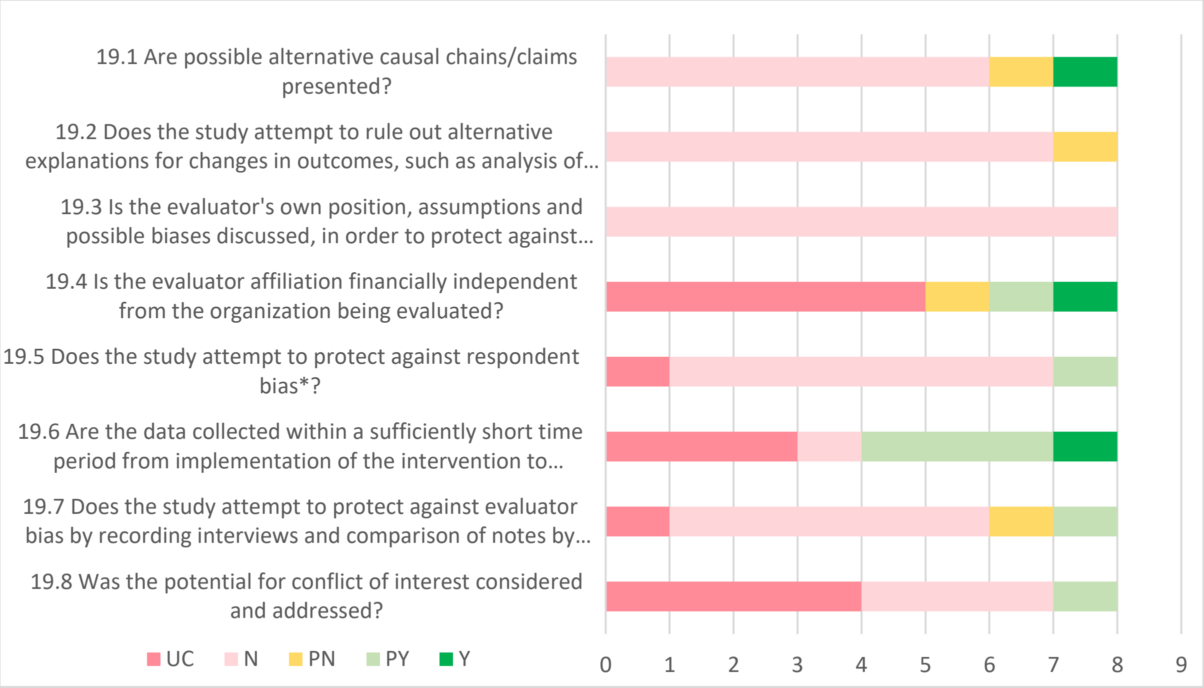
Criterion 17 “Sufficient independent information sources”



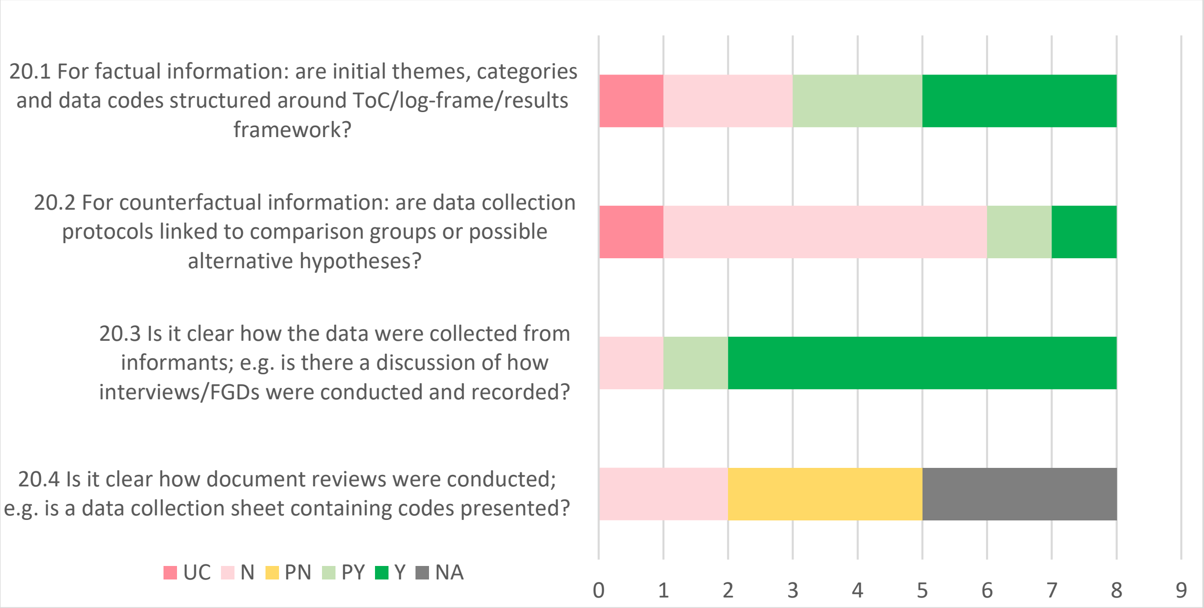
Criterion 18 “Triangulation of results from different information sources”



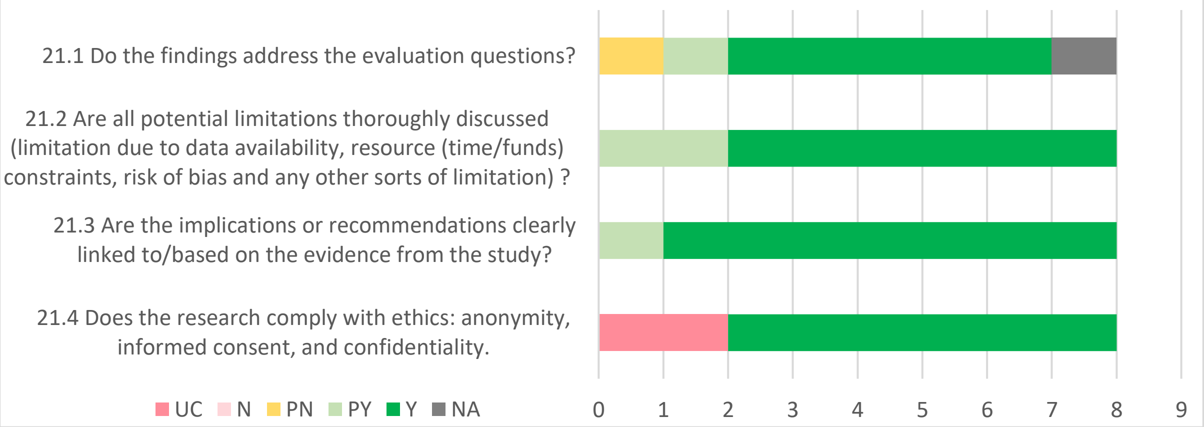
Criterion 19 “Discussion of bias”



Criterion 20 “Systematic, complete and transparent description of the data collection and analysis”



Criterion 21 “Discussion of the limitations of the evaluation”



Annex 6: Middle-level theories

MLT for Dialogue and Dissent

Middle-level theory (MLT)	(1) Developing CSO's capacity to deliver evidence-based L&A with clear definition of key issues increases their effectiveness	(2) Developing CSO's strong partnership with other CSO's and key stakeholder increases their effectiveness	(3) Support to CSO's L&A by creating political space leads to effective L&A activities	(4) CSO's activities with enhanced engagement of key actors leads to desired and impactful policy change and implementation	(5) CSO's L&A with mobilisation of community members/ local gate keepers results in desired community level change, and the mobilised community members get involved in L&A work
Enablers	Identification of good evidence	Inclusion of media	Engagement with different political parties (not just incumbent)	Use of (reliable) evidence-based L&A	Awareness of the issues & right among community members
	Clear content of the message	Engagement with citizen	Knowledge of personality (of politicians) and process of engagement	Adequate budget allocation for policy implementation	Mobilisation of local gate keepers
	CSO's skills and self-efficacy	Size of coalition		Involvement of key actors (within govt, private sector)	Safe space secured where vulnerable people can speak
	Participants of the training adopt approaches on which training was conducted	Inclusion of neutral actors (acceptable to government and private sector)		Transparency and monitoring of policy change and implementation	
	Better understanding of the needs of vulnerable people	Effective knowledge sharing among stakeholders		Awareness of government and donors (particularly	

				key actors like Prime Ministers) of the issues	
	Understanding of relevant national and/or international frameworks	Linking actors at different level (local and global etc.)		L&A activities are direct, rather than indirect	
	Use of systematic stakeholder mapping	Multi-sectoral approach		CSO's increased legitimacy and profile for L&A	
	Capacity development support is demand-driven			Inclusion of informal decision-making groups as lobby target	
	Availability of accurate and recent data for research publications and evidence production			Less confrontational L&A approach	
				Combination of top-down and bottom-up approach	
				Inter-ministerial approach	
Derailers	Mis-conceptualisation of key issues (e.g. "social inclusion”)	Lack of common working definition of key terms	Existence of ongoing or latent violent conflict	Fragmentation of government departments	Social norms that are against the changes
	Not internalising the training content within the organization	Lack of independence in media involved	Limited engagement due to pandemic or natural disaster	Lack of clarity with CSO on whom to talk to	If L&A activities use social media, limited use of social media by the targeted audience

	Evidence lacking hard data to convince government/private sectors	Lack of trust between CSOs	Fear of reprisal from criminals linked to private sector involved	Not creating relevant implementation body
		Interaction is not enough to build relationship and to better understand local context		Policy change is not known to the targeted population
				Limited operational mechanisms in place to implement change in policy
				Social norms that are against the changes
				Lack of cultural or contextual understanding of targeted actors (private sector) safeguard
Safeguards				Informal meetings with companies (without media) to build trust and understanding of company culture

MLT for Sexual and Reproductive Health and Rights

Middle-level theory	(1) L&A activities help improve rights and attitudes about SRHR for women, girls and disadvantaged groups	(2) Rights allow for SRHR choices for women, girls and disadvantaged groups	(3) Information about rights and services gives women, girls and other disadvantaged groups the knowledge to make informed choices	(4) Positive attitudes provide the supportive environment for realising SRHR	(5) Improved access to quality reproductive health services, helps promote their use, leading to improved SRHR outcomes
Enablers	There are appropriate fora for safe and regular discussion of rights and needs between government and vulnerable groups	People need to know their rights, for example the right to abortion or the right to refuse early marriage	Information about service availability allows people to make informed choices	Community gatekeepers such as NGOs, school staff, religious leaders, health workers, and government representatives support and raise awareness on key issues, such as the dangers of child marriage	For rights and information to lead to positive choices, the availability, accessibility and quality of SRHR services should be sufficient
	Decision making on policies, policy review processes, and policy implementation is inclusive of key groups such as women and informed by their needs	Governments develop or improve guidelines and action plans to enact policy in relation to harmful practices such as the sexual exploitation of children (SEC), FGM/C or child marriage	There are places where people, including vulnerable groups like sex workers, can obtain information, advice and support, and people know where to go to obtain this information	Public awareness of issues is increased, such as child marriage	Institutionalisation of training manuals, treatment protocols, and guidelines are available and used by health care staff that is sensitive to particular groups' needs, including appropriate care of vulnerable groups (e.g., people who use drugs)

Peer leaders and others from key communities are assisted in building skills (e.g., to express themselves clearly) and confidence (e.g., to speak in public), including for example through elected representatives, to participate in official discussion fora about SRHR programmes	Public service bodies recognise key groups e.g. trans-men and trans-women	Treatment literacy and mentoring programmes are available for key vulnerable groups including PLHIV	Acceptance about key issues, such as children's freedom of choice regarding marriage, and practices, such as checking young people's age before marriage, increase	Community health care staff are trained adequately, including in complementary areas such as mental health care (e.g., for people who use drugs or who are victims or survivors of sexual exploitation)
Partnerships with relevant groups and fora are made (e.g., to advocate for sex worker's rights as women's rights may require partnership with Women's Rights groups)	Rights are conditioned by the legal framework which would encompass areas such as legislation against/ laws criminalising harmful practices (e.g., child marriage, FGM/C) being enacted and enforced.	Appropriate groups are formed or trained to help key communities engage with rights or information about SRHR (e.g., youth clubs)	At-risk groups increase their engagement with their peers, e.g., children talking with friends and community members about child sexual exploitation	Community health centres with adequately trained staff are physically accessible to users
Partnerships with relevant groups and fora are made (e.g., to advocate for sex worker's rights as women's rights, partnership with Women's Rights groups)	Police and judicial officers investigate allegations of sexual exploitation, leading to the arrests of perpetrators	Key groups have knowledge about protective laws (e.g. girls know about legal age of marriage/laws against child marriage, FGM/C)	Community leaders initiate discussions within their communities on change of values, norms and practices (e.g. keeping children safe from sexual exploitation)	Monitoring systems to assess access and quality of services, including for key vulnerable groups, are established and used by decision makers

Appropriately trained groups are established to support access of vulnerable groups to justice, such as community paralegals to raise awareness and work with perpetrators and victims, to reduce violence and support LGBT people to monitor and report acts of violence against them	Law enforcement agencies apply child-friendly protocols	Key groups have knowledge about diseases and services (e.g., HIV and modern contraception)	The private sector effectively implements, and monitors within their sector, relevant memoranda of understanding for child rights safeguarding, including the protection against and reporting of child sexual exploitation	There is a safe and convenient environment for access to services for women and vulnerable groups. Gender sensitive services are available, including female peer educators, women only service hours, to facilitate and support networks of women. Other vulnerable groups such as sex workers and those who use drugs have access to high-quality STI services at a convenient time and in a safe place. Access may be through fixed government and NGO services or mobile units.
CSOs and the public enter into dialogue with target industry groups regarding the prevention and reporting of harmful practices (e.g. sexual exploitation of children).	Victims and survivors are compensated after identification and litigation of crimes (e.g. regarding sexual exploitation of children or human trafficking)		Decision making power increases, such as girls' decisions about whether to stay in school or whether or not to get married	Supply and availability of medication is responsive to the needs of key groups, including vulnerable groups like the young and adult gay people, men who have sex with men and transwomen communities.

	Peer leaders and others from key communities are assisted in building skills and confidence (e.g., speaking in public, to express themselves clearly) to participate in L&A in SRHR programmes	Relevant bodies like branches of service-delivery NGOs receive official registration to gain access to formal government mechanisms and complementary government services			Institutionalisation of training manuals, treatment protocols, and guidance that is sensitive to particular groups' needs
		Community-based child protection mechanism and referral systems for victims and survivors of child sexual exploitation are in place and effective			Increased access to youth-friendly health facilities
					Service users feel confident and supported in consulting appropriate sources about SRHR issues
Derailers	Lack of anti-discrimination legislation or its promotion	Negative backlash by media and cyber-bullying towards LGBT following attempts to legitimise consensual same-sex intimacy between adults	Knowledge about legality of harmful practices may lead to them being done in secret (e.g., FGM/C, child marriage) or camouflaged (e.g., FGM/C alongside male circumcision)	Regressive attitudes of community members and gatekeepers towards key issues, such as teenage pregnancy or the belief that marrying off pregnant girls is the best solution	Promoting the use of modern contraceptives will not be effective if they are not available, which can be the case especially in rural areas in many developing countries.

	Lack of common language among advocacy groups	Increased mobilisation of the sex worker community has led to increased negative media attention on sex work and sex workers		Adverse reactions of non-targeted groups (e.g., jealousy), such as boys and men in the case of programmes for women and girls	If behaviours are already adopted, programmes to promote the use of modern contraceptives will have little impact if they are already widely adopted
	Attitudes of community members (e.g., son preference) reduce efficacy of messaging.	Sexual harassment and aggression in public places or by public officials (e.g., schoolteachers)		Some environmental factors can work against realisation of rights such as economic situation of the household, pregnancy and age (where youths are simply unable to meaningfully participate in community decision-making)	
		Promoting demand for contraceptives will be less effective if women – or their relatives – want many children, so the appropriate intervention may need to incorporate messaging around family size, or tackling the factors that make large families attractive, such as high child mortality and son preference		Women may be constrained in their use of contraceptives by their partners or other family members.	

Safeguards	Harmful traditional practices prevention committees are established by the government and strengthened by community organisations to enable practitioners to stop harmful practices in their communities. These might include mobilisation of health extension workers to monitor and supervise the prevention of these practices.	Communication around SRHR can be sensitive, and so finding appropriate channels through which men and women will meaningfully engage should take into account local norms and values. And health workers need the communication skills to apply this approach.	Relevant gatekeepers need to be the targets of behaviour change communication, whether these are community and faith leaders, fathers, mothers or mothers-in-law, husbands, and boys.	Ensure service providers have insufficient resources (e.g., contraceptives) to provide for targeted groups adequately

Annex 7: Terms of Reference – Evaluating Lobbying and Advocacy: an assessment of 32 partnership evaluations

Introduction and Rationale

These Terms of Reference (ToR) present the outline for a study on the evaluations of the different strategic partnerships of the Dialogue and Dissent (D&D) programme (2016-2020) and the Sexual and Reproductive Health and Rights (SRHR) Partnership Fund (2016-2020).

One of the functions of the Policy and Operations Evaluation Department (IOB) is to advise the policy departments at the Ministry of Foreign Affairs (MFA) and their partners about evaluation quality. On several occasions, Civil Society Organisations (CSOs) and various policy departments at the MFA have requested IOB to provide more guidance on evaluations and evaluation methods that can be validly used to evaluate programmes on lobbying and advocacy. The final evaluations of the D&D programme and the SRHR Partnership Fund provide a great opportunity to do this. The Strategic Partnerships (SPs) supported through these programmes have recently ended and the mandatory external end evaluations have all been submitted in 2021.

This study will assess the evaluation methodologies of the 32 external end evaluations: 25 for the D&D programme and 7 for the SRHR Partnership Fund. Based on these assessments, the study will formulate lessons and recommendations regarding evaluating lobby and advocacy programmes. It will expand on the general framework for attribution of cause and effect in small n impact evaluations, as put forward by White and Phillips (2012), especially for evaluations on lobbying and advocacy.

The findings and lessons of this exercise will be relevant for the various policy departments at the MFA, (I)NGOs and CSOs in subsequent policy frameworks (e.g., in Power of Voices) and for the evaluators of the partnerships.

Where possible, the study will also separately describe the results achieved by the 32 separate partnerships. It will not be possible to draw conclusions about the effectiveness of both programmes as a whole, because additional primary research will be necessary. In 2024, therefore, IOB will perform a broader evaluation of Dutch policy on civil society strengthening. That evaluation will investigate whether long term effects have been achieved by the subsequent policy frameworks supporting strategic partnerships.

Background

Since The Netherlands' government engaged in development cooperation, the relationship between the Ministry of Foreign Affairs and CSOs has taken on many different forms. From 2013 onwards, the MFA envisaged a more political role for CSOs, in reinforcing civil society dialogues between citizens, government and the private sector (IOB, 2019). The policy framework 'Dialogue and Dissent' built on these recommendations and set out the principles for CSOs to enter into a strategic partnership in the area of 'lobbying and advocacy' with the Ministry in the 2016-2020 period.[9]

[9] Applying (consortia of) CSOs were requested to show a track record on lobby and advocacy and experience in strengthening civil society in low- and lower-middle income countries and submit a Theory of Change (ToC). After the selection, the partners and the ministry would jointly formulate a strategic goal and envisaged results and only after that, the partners were asked to draw a programme proposal.

The main objective of the Dutch policy was to strengthen CSOs in low- and lower-middle income countries in their roles as advocates and lobbyists. This role was seen as essential for holding policymakers, government and private sector to account, and as a way for CSOs to contribute to inclusive economic growth and development and to help reduce inequality.

There was no actual thematic delimitation for the consortia that wanted to apply for funds in the Dialogue and Dissent programme, as long as objectives were in some way connected to the policy agenda as set out in the policy document 'A World to Gain'. Potential partners were free to address any of the issues identified in that document. The four main policy priorities of Dutch development cooperation were (i) women's rights and sexual and reproductive health and rights (SRHR); (ii) water; (iii) food security; and (iv) security and rule of law.

In addition to Dialogue and Dissent, the Dutch MFA also introduced the SRHR Partnership Fund for the same period 2016-2020. Organisations were allowed to submit proposals for both Dialogue and Dissent and for the SRHR Partnership Fund. The strategic partnership modality and the tendering procedure were similar. A difference between the two was that for Dialogue and Dissent, partners were not allowed to provide service delivery to beneficiaries, while providing access to SRHR services was one of the objectives of the SRHR Partnership Fund.

The MFA financed 25 partnerships through D&D and 7 via the SRHR Partnership Fund (see more details in Annex 1). In total, the ministry committed EUR 1.14 billion for these programmes for the period 2016-2020. Activities funded from these two programmes were implemented in 129 countries.

The grant agreements for the 32 partnerships prescribed that independent final evaluations of the effects of the programme had to be conducted. The ToR and methodology of these evaluation had to be approved by a mutually agreed upon independent external advisory group. In addition, the final report had to comply with the quality standards for external evaluation, as set out in the IOB guidelines that were attached to the grant decisions (see Annex 2).

Recently, IOB updated its evaluation quality criteria (see Annex 3). The content and tendency of the updated list of 26 evaluation criteria is the same as the guidelines provided. The renewed quality criteria are more user friendly. The main difference is that the updated version provides more guidance on what is sufficient and illustrates the application of the criteria with various examples.

Objectives and delimitation

This study has the following objectives:

- It will assess the methodology of the 32 evaluations carried out under the D&D programme and SRHR Partnership Fund programme;
- It will formulate lessons and recommendations with regards to evaluating lobby and advocacy programmes;
- Where possible, it will describe the results achieved by the 32 programmes;
- This study will systematically assess if the used methods as applied in the evaluations were appropriate to answer the respective research questions. It is important to note that this exercise only focusses on the evaluation reports and the deployed methods and does not assess the entire evaluation process. The focus of the assessment will be on research questions regarding the OECD/DAC criteria for effectiveness and, where possible, impact. The synthesis will explicitly not focus on research questions regarding the efficiency, relevance, coherence and sustainability of the programmes.

- The study should provide practical guidance for evaluating lobby and advocacy programmes. By assessing the evaluative material produced for the two partnership programmes this study should shed light on how to address research questions on causality for lobby and advocacy programmes.
- Where possible, the study will also separately describe the results achieved by the 32 programmes, based on the evaluation findings and the assessment of the methodologies. Described results can include (i) the strengthened role of CSOs in their roles as advocates and lobbyists and (ii) thematic results, such as on SRHR, climate or food security.[10]

Research questions

The following five research questions will guide the study.

1. Which evaluation methodologies are used in the 32 evaluation reports to answer the research questions on effectiveness and, when available, impact? Does the way in which the methodology are applied in the evaluation reports correspond to the methodology in theory?
2. Are the evaluation methodologies as applied in the 32 reports in line with the updated IOB evaluation criteria that focus on evaluation methodology (criteria 10-21 see Annex 3 and section 5)?
3. What are the appropriate evaluation methods, and their common characteristics, for evaluating effectiveness, in the field of capacity building of CSOs for lobby and advocacy?
4. What were the common characteristics for the less suitable methods to evaluate capacity building of CSOs for lobby and advocacy?
5. Based on the evaluation reports and the assessment of the evaluation methodologies, what can be said about the achieved results of the 32 supported partnerships?

Note that the methodology as mentioned in an evaluation report may not fully correspond to the followed methodology in the evaluation. It is therefore important to distinguish the methodology in theory (perfectly applied) from the methodology as actually applied in the evaluations.

Proposed methodology

The assessment of the evaluation methodologies will be done on the basis of IOB's evaluation quality criteria. Specifically, the following 11 criteria that focus on evaluation methodology can provide a framework.[11]

- The research design is clearly elaborated and shows how the research results will contribute to answers to the evaluation questions
- The methods are appropriate to evaluate effectiveness: attribution and / or contribution (if effectiveness is an evaluation criterion/question)
- The indicators or result areas are appropriate to capture the planned results along the different levels in the ToC
- Justified choice of sample, cases and information sources (e.g. choice of countries, projects, organisations and persons)

[10] The description of results is not a main objective of this study; the extent to which this is possible depends on the evaluative quality of the underlying material. It will not be possible to draw conclusions about the effectiveness of both programmes based solely on the 32 evaluation reports, because primary research is necessary to take synergies and coherence between the programmes into account.
[11] Note that criterion 12 has been excluded because this study does not focus on evaluating efficiency.

- The analyses are appropriate, given the chosen research design
- Summary of the methodology in an evaluation matrix
- Sufficient independent information sources
- Triangulation of results from different information sources
- Discussion of bias
- Systematic, complete and transparent description of the data collection and analysis
- Discussion of the limitations of the evaluation

We suggest that two persons first independently assess the 32 evaluations, reporting a judgement (good, sufficient, insufficient) accompanied by arguments, then come to a consensus.

The identification of appropriate evaluation methods, and their common characteristics follows the exercise presented above. This section should build on the general framework for attribution of cause and effect in small n impact evaluations, as put forward by White and Phillips (2012). That paper formulated a general framework for qualitative evaluation methods using four methods that can make a plausible claim for effectiveness.[12] [13]

The consultant must also consider new insights gained over the last 8 years, and consider the particularities of interventions strengthening CSOs for lobby and advocacy. The exercise should result in a (possibly revised of refined) set of common characteristics of appropriate research methodologies for lobby and advocacy programmes. The consultant should be explicit about the approach for causal inference that would be useful to apply.

The description of results must take the strength and validity of the evidence into account; results that depend on inappropriate evaluation methods should not be included. When describing the results, IOB proposes to focus on (i) the strengthened role of CSOs in their roles as advocates and lobbyists and (ii) thematic results, such as on SRHR, climate or food security.

The methodology proposed here can be further refined in agreement between IOB and the consultant.

Organisation

The assignment is contracted by the Policy and Operations Evaluation Department (IOB) of the Netherlands Ministry of Foreign Affairs. Within IOB, Ferko Bodnár and Caspar Lobbrecht are responsible for the organisation and management of the implementation of the study. More specifically, they will:

- Contract a team of researchers/consultants for conducting the review. This contract will include specific milestones for the delivery of outputs.
- Provide the consultant(s) with all necessary documentation, including the 32 programme documents, Terms of References for the external evaluations and the final evaluation reports.
- Supervise the implementation and progress of the review through regular (virtual) meetings with the consultant(s).

[12] These were realist evaluation, contribution analysis, process tracing and general elimination methodology.

[13] Another useful source is the Vaessen et al. (2020) guidebook for evaluators. This book presents the main features and procedural steps, advantages and disadvantages for the most common evaluation methods used in international development.

- Arrange for internal and external quality control of the review process and its outcomes in line with IOB requirements.

The assignment is subject to IOB's regular quality control system. The internal reference group consists of Rob van Poelje (chair), Kirsten Lucas and Jelmer Kamstra.

The external reference group consists of Rob van Poelje (chair), Ronald Siebes (MFA-DMM, previously MFA-DSO/MO), Cobi Mars (MFA-DSO MEL), Ini Huijts (MFA-DSO/GA), Karen Biesbrouck (OXFAM), Karel Chambille (HIVOS), Jos Vaessen (IEG).

The external reference group has an advisory role and will be asked to provide feedback on:

- ToR and the adjusted ToR;
- the inception report;
- the draft report;
- the final report.

For the inception, draft and final report, IOB will organise feedback sessions in which the reports can be discussed with the (team of) independent consultant(s).

Consultant

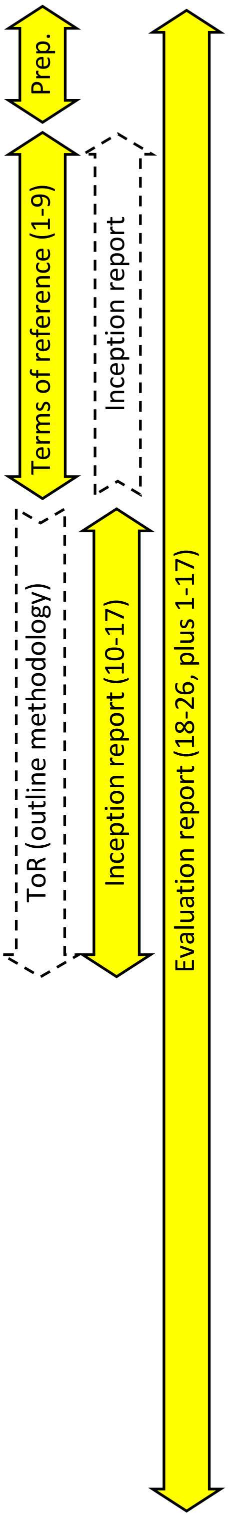
IOB will subcontract the undertaking of the study to a (team of) independent consultant(s) though a direct award.

Deliverables

- Inception report. This report should present the fine-tuned proposed methodology and approach.
- Draft report for RQ 1-4, no longer than 30 A4 and accompanied by an executive summary no longer than 4 A4.
- This excludes the assessments the individual evaluations – these may be presented in a separate Annex.
- Draft report for RQ 5
- Final report for RQ 1-4 and final report for RQ 5.

Based on the draft reports, IOB will organise a discussion session with the consultants and the involved partners of the 32 evaluations.

Updated IOB evaluation criteria



Quality control of the evaluation

1. A reference group oversees the evaluation
2. Evaluators are independent *

Description and background of the intervention

3. Description of the context of the intervention
4. Description of the intervention *
5. Validation of the assumptions underpinning the ToC or result chain*

Objective and delimitation of the evaluation

6. Description of the objective of the evaluation
7. Delimitation of the evaluation

Evaluation questions

8. Choice of OECD-DAC evaluation criteria to be covered
9. Clear set of evaluation questions

Evaluation methodology

10. The research design is clearly elaborated and shows how the research results will contribute to answers to the evaluation questions *
11. The methods are appropriate to evaluate effectiveness: attribution and / or contribution (if effectiveness is an evaluation criterion/question) *
12. The methods are appropriate to evaluate efficiency (if this is an evaluation criterion/question)
13. The indicators or result areas are appropriate to capture the planned results along the different levels in the ToC *
14. Justified choice of sample, cases and information sources (e.g. choice of countries, projects, organisations and persons) *
15. The analyses are appropriate, given the chosen research design *
16. Summary of the methodology in an evaluation matrix
17. Sufficient independent information sources *
18. Triangulation of results from different information sources
19. Discussion of bias
20. Systematic, complete and transparent description of the data collection and analysis *
21. Discussion of the limitations of the evaluation *

Results and conclusions

22. Conclusions answer research questions *
23. Conclusions follow logically from the research findings *
24. Validation of draft conclusions

Usefulness and readability of the evaluation report

25. Recommendations should be useful and practical, given the evaluation objectives and its intended users
26. The report is well readable, consistent, and includes a clear summary with evaluation objective, evaluation questions, conclusions and recommendations