**IOB EVALUATION CRITERIA 2024**

# **Introduction**

This document presents IOB’s revised and updated (2024) evaluation quality criteria. The 17 criteria serve as a comprehensive framework for assessing and monitoring the quality of project and programme-level evaluations commissioned by the Ministry of Foreign Affairs or by its partner organizations. In addition, IOB uses the evaluation quality criteria in its own research.

IOB first published its initial evaluation quality criteria in 2012. The purpose and most underlying principles behind IOB's criteria have remained relatively constant since then. Initially, IOB employed a list of twenty evaluation quality criteria, categorized under validity, reliability, and usefulness. In 2021, IOB expanded this list, offering more in-depth explanations on interpretation and application, along with illustrative examples.

In 2022, IOB commissioned a [review of evaluation methodologies](https://english.iob-evaluatie.nl/publications/reports/2023/04/15/effectiveness-support-to-lobby-and-advocacy), specifically focusing on methods used to evaluate lobby and advocacy related activities and projects. The findings of that review, published in 2023, have been incorporated into this updated version of IOB’s evaluation quality criteria. As a result, this document presents the seventeen updated criteria, accompanied by an explanation as to why IOB deems an evaluation practice as ‘good’, ‘adequate’, or ‘inadequate’, along with illustrative examples.

The IOB evaluation criteria have been structured chronologically across three different phases in the evaluation process: the Terms of Reference, the elaborated methodology (e.g. in an ‘inception report’ or ‘technical proposal’), and the (draft and final) evaluation report – see graph 1.

*Graph 1. Evaluation criteria per evaluation phase*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | Phase I  Terms of Reference | Phase II  Elaborated methodology | Phase III  Draft and final report |
| 1. Reference group | | |  |  |  |
| 1. Independence of evaluators | | |  |  |  |
| 1. Description of context | | |  |  |  |
| 1. Description of the intervention | | |  |  |  |
| 1. Scope of the evaluation | | |  |  |  |
| 1. Evaluation’s objective | | |  |  |  |
| 1. OECD/DAC evaluation criteria | | |  |  |  |
| 1. Evaluation questions | | |  |  |  |
| 1. Research design | | |  |  |  |
| 1. Plausibility of causal claim | | |  |  |  |
| 1. Indicators or description of result areas | | |  |  |  |
| 1. Sampling and case selection | | |  |  |  |
| 1. Sufficient and external information sources | | |  |  |  |
| 1. Description of limitations and bias | | |  |  |  |
| 1. Transparency | | |  |  |  |
| 1. Answer to evaluation questions | | |  |  |  |
| 1. Logic of conclusions and recommendations | | |  |  |  |
|  | Main assessment |
|  | Re-assessment |

1. During formulating the Terms of Reference, the first subset of criteria (1 – 8) help shape aspects such as project description, evaluation objectives, scope, evaluation questions, and quality control. The ToR should broadly present and assess the intervention theory and assumptions of the project under evaluation as this will provide useful input for formulating the evaluation questions.
2. The second subset of criteria (9 – 14) help in assessing the elaborated methodology, as outlined by the evaluator (e.g. in an ‘inception report’ or ‘technical proposal’). Based on the assessment of these criteria, a commissioner may ask evaluators to adjust the methodology. This phase also includes a re-assessmentof the first subset of criteria, to make sure whether the elaborated methodology is in line with the ToR.
3. The third subset of criteria (15 – 17) are designed to assess the quality of the draft and final report. These criteria focus on transparent reporting and on conclusions and recommendations formulated. During this stage of assessing evaluation quality, it is no longer possible to adjust the data collection as applied, but it remains possible to improve descriptions, enhance the analysis, and to reformulate conclusions. This phase includes a re-assessment the second subset and of five criteria from the first subset.

Section 2 of this document presents the one-page summary of the IOB evaluation criteria. Section 3, then, delves deeper into the updated evaluation criteria, and outlines the conditions that determine whether a criterion can be assessed as ‘good’, ‘adequate’, or ‘inadequate’. Each criterion contains a clickable (and expandable) assessment grid, providing detailed insights into the specific conditions that determine whether a criterion can be assessed as ‘good’, ‘adequate’, or ‘inadequate’. Note that several criteria contain sub criteria (a, b, c), and each of these must at least be assessed as adequate. This document also uses various clickable (and expandable) examples from evaluations of development cooperation and from broader foreign policy and diplomacy to illustrate how the criteria should be applied in practise. Annexes 1, 2 and 3, present the joint assessment grids.

The intended target group of this document include:

* M&E staff of partner organizations responsible for commissioning, managing, or evaluating the quality of external evaluations;
* Policy officers at the ministry overseeing the commissioning, management, or assessment of external evaluations;
* Researchers responsible for conducting evaluations of projects or programmes financed by the ministry, and finally;[[1]](#footnote-2)
* IOB researchers.

These IOB evaluation quality criteria are relevant for ex-post evaluations that aim to account for expenditure and focus on effectiveness. In case effectiveness is also part of a MTR assignment, these criteria can also be applied to MTRs. The criteria are less applicable to ex-ante evaluations, scenario analyses, feasibility studies or studies that focus entirely on learning.

The criteria are equally relevant for development cooperation as they are for other realms of foreign policy, such as diplomacy or lobby and advocacy.

IOB recognises that the art and craft of policy evaluation is constantly evolving: we welcome feedback and suggestions via [iob@minbuza.nl](mailto:iob@minbuza.nl).

# **One-page summary**

|  |  |  |
| --- | --- | --- |
| **Phase I – Terms of Reference** | | |
| 1 | Reference group | The ToR presents a reference group, offering guidance on different phases of the evaluation process. |
| 2 | Independence of evaluators | The evaluators should not have been involved in the design or implementation of the project and maintain complete impartiality. |
| 3 | Description of context | The ToR describes the context in which the project has been implemented and the rationale behind it. |
| 4 | Description of the intervention | The ToR presents the intervention theory, or Theory of Change, of the project evaluated. |
| 5 | Scope of the evaluation | The ToR mentions the period, countries, and expenditure evaluated. |
| 6 | Evaluation’s objective | The ToR articulates the evaluation's objectives and the intended utilization. |
| 7 | OECD/DAC evaluation criteria | The ToR considers which OECD/DAC criteria and other cross-cutting topics are included in the evaluation. |
| 8 | Evaluation questions | The ToR should formulate precise and testable evaluation question that align with criteria 5, 6 and 7. |
| **Phase II – Elaborated methodology** | | |
| 9 | Research design | The research design encompasses the overall approach and evaluation methods, appropriate to validly and reliably answer the evaluation questions and achieve the evaluation’s objective. It should enable triangulation of findings. |
| 10 | Plausibility of causal claim | The selected evaluation methods (either qualitative or quantitative) must be appropriate to evaluate the contribution or attribution of the projects to results at the outcome or impact levels. |
| 11 | Indicators or description of result areas | The evaluators should propose valid quantitative indicators, or qualitatively describe the result areas as concretely as possible. |
| 12 | Sampling and case selection | The report provides a valid justification for all sampling and case selection strategies, and is transparent about criteria applied. In quantitative research, sample size is based on a power calculation. |
| 13 | Sufficient and independent information sources | The proposed research design should present which information sources will be included in a way that minimizes source bias, by using sufficient and independent information sources. |
| 14 | Description of limitations and bias | The 'inception report' or 'technical proposal' should clearly describe the limitations in the reliability and validity (both internal and external) of the proposed methodology, as well as any potential biases, including selection bias. |
| **Phase III – Draft and final report** | | |
| 15 | Transparency | The evaluation report should present all research methods, data collection methods, data sources, and data analysis techniques employed in a transparent and complete manner |
| 16 | Answer to evaluation questions | The draft and final report should provide an answer to all evaluation questions. In case of unforeseen limitations, as a minimum the report must highlight which questions it did not answer and explain why. |
| 17 | Logic of conclusions and recommendations | There should be a clear and rational progression in the evaluation report from findings to conclusions, and from the conclusions to the recommendations. Each conclusion should be supported by specific findings, and the implications of limitations and bias are sufficiently taken into account. |

# **Phase I – Terms of Reference**

## **Reference group**

The reference group plays a crucial role in guaranteeing the quality and independence of the evaluation process. It serves as an advisory body to the evaluation commissioner, offering guidance on the Terms of Reference (ToR), the selection of evaluators, the detailed methodology section (in the inception report), and the draft evaluation report.

The reference group typically comprises the evaluation commissioner, a representative from an implementing partner organization, and individuals with experience in relevant themes and evaluation methodologies. It should include at least one but preferably two or more independent member(s). To ensure downward accountability and sufficiently take the context into account, IOB recommends including (a) reference group member(s) from target countries.

### 🡸 **Click here to view the assessment grid for criterion 1**

|  |  |  |
| --- | --- | --- |
| 1.a | The ToR presents a reference group with two or more independent and external members | **√√√ Good** |
| The ToR presents a reference group with one independent and external member | **√√ Adequate** |
| The ToR presents a reference group without independent and external members | **ꓫ Inadequate** |
| 1.b | The ToR requests that the reference group will be involved in different stages of the evaluation (TOR, inception report, draft final report) | **√√√ Good** |
| The ToR does not propose that the reference group will be involved in the different stages of the evaluation (ToR, inception report, draft final report). | **ꓫ Inadequate** |

## **Independence of evaluators**

The evaluators and affiliated organisations should not have been involved in the design or implementation of the project under evaluation – either with the organisation responsible for implementation or at the MFA. Additionally, they must maintain complete impartiality with no vested interest in the ultimate outcome of the evaluation.

### 🡸 **Click here to view the assessment grid for criterion 2**

|  |  |  |
| --- | --- | --- |
| 2 | The ToR requests that the evaluators should not possess involvement in the design or implementation of the project | **√√√ Good** |
| If the ToR does not mention that researchers should be independent | **ꓫ Inadequate** |

## **Description of context**

The ToR should present an adequate description of the national, sectoral, and political context in which the project or programme has been implemented, and also mention the rationale and problem analysis of the project. If available, the ToR should also present relevant baseline data on project and other relevant result indicators. Note that benchmarks for assessing progress, such as baseline date, do not have to be exclusively quantitative and can also encompass qualitative data.

### 🡸 **Click here to view the assessment grid for criterion 3**

|  |  |  |
| --- | --- | --- |
| 3.a | The ToR presents baseline data about project result indicators | **√√√ Good** |
| The ToR does not present baseline data but includes an explicit strategy to assess progress and to (re)construct the baseline situation. | **√√ Adequate** |
| There are no baseline figures on project result indicators and there is no strategy to assess progress or to (re)construct a baseline situation. | **ꓫ Inadequate** |
| 3.b | The ToR presents a clear context and problem analysis which forms the rationale for the project’s interventions. | **√√√ Good** |
| There is no (or insufficient) problem and context analysis | **ꓫ Inadequate** |

## **Description of intervention**

The ToR should present a specification of a project’s intervention logic, ‘policy theory’, or Theory of Change (ToC).[[2]](#footnote-3) When there is no intervention logic or ‘policy theory’, the evaluator may need to reconstruct one by using project documentation.[[3]](#footnote-4)

The ToR should describe and illustrate of how and why a desired change is expected to happen in a specific context: the description of the ToC should articulate the implicit and explicit assumptions underpinning the relationships in the results chain (e.g. between output and outcome) and may assess the existing evidence base evidence for these assumptions and links. Herein, it is essential to describe ‘the missing middle’ between a project and the desired outcome and impact. In addition, the (reconstructed) ToC should consider contextual external factors that may, as ‘rival explanations’, influence results at various levels in the results chain (input, activity, output, (intermediate) outcome, impact).

### **🡸 Click here to view the assessment grid criterion 4**

|  |  |  |
| --- | --- | --- |
| 4.a | The ToR presents an intervention logic, theory or ToC and addresses intermediate steps between inputs, activity, output, outcome, and impact. | **√√√ Good** |
| The description of the intervention distinguishes between activities, outputs and outcomes. The results chain should make sense and not omit important steps. | **√√ Adequate** |
| There is no specified or reconstructed description of the intervention, or it directly links implemented activities to results at outcome or impact level without distinguishing intermediate results. | **ꓫ Inadequate** |
| 4.b | The ToR considers external factors that may influence outcomes at various levels in the results chain | **√√√ Good** |
| The ToR does not mention any external factors that can influence outcomes at various levels in the results chain | **ꓫ Inadequate** |
| 4.c | The ToR explicitly assumptions underlying the presented relationships at different levels in the results chain | **√√√ Good** |
| The specified or reconstructed ToC does not mention any assumptions | **ꓫ Inadequate** |

### **🡸 Click here to read examples for assessing criterion 4**

|  |  |  |
| --- | --- | --- |
| **Example ‘good’**   |  | | --- | | 4.a: Description of the intervention theory:   * A coalition of Dutch NGOs aims to influence decision-makers' agendas, shape political debate, and contribute to policy and legislative reforms to promote social change and reduce inequalities (impact). Achieving this requires strengthening civil society organizations and building coalitions with like-minded actors across various sectors and regions in a sustainable manner (outcome). To support this, Dutch NGOs focus on increasing engagement between decision-makers and civil society, fostering productive policy dialogues among CSOs, government bodies, and societal actors. They also seek to raise public support and media attention for the issues highlighted by their advocacy efforts (intermediate outcome). Key outputs include the number of public awareness campaigns organized, policy briefs produced, trainings conducted to enhance the lobbying and advocacy capacities of CSOs, and grassroots CSO networks established. Activities encompass designing and implementing public awareness campaigns on key social issues, conducting research and analysis on pressing social issues and policy gaps, providing capacity-building workshops and training sessions for local CSOs, and connecting and strengthening networks across grassroots CSOs.   4.b. Description of external factors:   * Civic space and the political will and openness of local and national governments to engage in dialogue with civil society actors. * The legal and social environment in the country, including the level of civil and political freedoms. * The influence and role of international institutions and other donor agencies in shaping local policy contexts. * The strength of local CSOs and their ability to organize and mobilize marginalized communities. * The perception of international NGOs and local partners by the target decision-makers and stakeholders.   4.c Description of assumptions:   * If local governments are open to engaging with civil society, then lobbying and advocacy efforts will be more effective. * If Dutch NGOs and local CSOs align their priorities, they are more likely to create a unified front that can influence policy change. * If marginalized groups can effectively mobilize and express their needs, then their voices will be heard in political debates and decision-making processes. * Policy reform is influenced by both national and international advocacy, meaning that multilevel advocacy is crucial for sustainable impact. * The collaboration between civil society and governmental actors leads to policy changes that address the root causes of exclusion and marginalization. * Civil society organisations have insufficient policy influence because they lack the capacity to effectively conduct lobby and advocacy. * Training provided to CSOs will sustainably enhance their ability to effectively engage in lobbying and advocacy. |   **Example ‘inadequate’**   |  | | --- | | 4.a: Description of the intervention theory:  With the goal of promoting social change and reducing inequalities, Dutch NGOs organise public awareness campaigns and provide capacity-building workshops and training sessions for local CSOs  4.b: Description of external factors   * Absent   4.c: Main assumptions   * Absent | |

## **Scope of the evaluation**

The ToR should clearly delineate the evaluation period, geographical focus and specific policy domain (programs, projects) and expenditure that will be a part of the evaluation. Moreover, it should also specify which particular components of the intervention theory or ToC are to be included in the evaluation.

### **🡸 Click here to view the assessment grid criterion 5**

|  |  |  |
| --- | --- | --- |
| 5 | The ToR clearly defines the evaluation period, the geographical focus and the expenditure included in the evaluation | **√√√ Good** |
| The ToR does not clearly define the evaluation period, the geographical focus or the expenditure included in the evaluation | **ꓫ Inadequate** |

## **Evaluation’s objective**

The ToR should distinctly articulate the evaluation's objectives and the intended utilization of its results. The ToR should explain who are the intended users of the results of the evaluation and for what purpose the results will be used. Typically an evaluation informs policymakers, funders, and (parliament’s) policy and financial oversight organizations, but it could also inform other stakeholders, including implementing agencies and final beneficiaries.

Ensuring clarity in the ToR about the specific objectives can shape the research questions and the type of recommendations required. Evaluations often serve multiple objectives, which can be categorized into:

1. **Knowledge objectives**: These involve understanding whether the stated objectives have been achieved, contributing to accountability. It also seeks to uncover lessons about what works, why, and how, thereby contributing to learning.
2. **Action objectives**: These aim at providing recommendations. An action objective can determine when evaluation results are needed, for example for subsequent project phases or formulation of new policy.

### **🡸 Click here to view the assessment grid criterion 6**

|  |  |  |
| --- | --- | --- |
| 6 | The ToR articulate a distinction between the knowledge and action objectives in the evaluation. The objectives are stated, and it is evident who will use the evaluation results and for what purpose. | **√√√ Good** |
| The ToR has a clear knowledge objective, but it remains unclear how the findings will be utilised. | **√√ Adequate** |
| The ToR solely mentions an action objective or aims to showcase effectiveness rather than evaluating it. | **ꓫ Inadequate** |

### **🡸 Click here to read examples for assessing criterion 6**

|  |
| --- |
| Example good   * Knowledge objective: ‘*The primary aim of this evaluation is to assess the effectiveness of Project X in achieving objectives Y and Z, after five years of implementation and provide a comprehensive understanding of the factors contributing to this effectiveness.’* * Action objective: ‘*The insights gained from evaluating the effectiveness of Project X and understanding the underlying factors will ensure that policymakers can make informed decisions. This includes deciding whether to continue the project and, if continued, identifying specific areas for improvement.’*   Example adequate   * There is no distinction between knowledge objective and action objective: *‘The primary aim of this evaluation is to assess the effectiveness of Project X after five years of implementation and provide a comprehensive understanding of the factors contributing to this effectiveness.’*   Example inadequate   * Unclear: *‘The objective of this evaluation is to ascertain the contribution of project X in a changing environment.’* (Contribution to what? Which changes in environment?) * The objective is to showcase effectiveness: *‘The objective of this evaluation is to demonstrate that project X has been effective, with the aim of securing funding for a second phase.’* * The objective of the evaluation is to see how a specific project can be improved, without directing the evaluation in any way. |

## **Selection of OECD/DAC criteria and cross-cutting topics**

The ToR should carefully consider the selection of relevant [OECD/DAC Evaluation Criteria](https://www.oecd-ilibrary.org/development/applying-evaluation-criteria-thoughtfully_543e84ed-en;jsessionid=Yyykvok1XILTuI5TlqLFjCOci2EUXmuGyp7-h4Fh.ip-10-240-5-71) and other cross-cutting topics.

* Given the specific evaluation objectives (criterion 6) and scope (criterion 7), not all OECD/DAC evaluation criteria (relevance, coherence, effectiveness, efficiency, impact, sustainability) may be applicable. The ToR should include an assessment of which OECD/DAC evaluation criteria will be integrated into the evaluation questions.
* The ToR may also incorporate other relevant cross-cutting issues (e.g. gender, localisation, poverty reduction, or climate), accompanied by a rationale.

In order to optimize the use the value of the OECD/DAC criteria ‘relevance’ and ‘coherence’ in evaluation research, the criteria are put in a logical order: relevance, coherence, effectiveness, efficiency, sustainability and impact.

For the ‘relevance’ criterium, two dimensions are central: (1) the extent to which policy objectives, design and implementation of policy interventions are in line with the needs of the target group and other non- policy makers stakeholders and (2) the extent to which the policy intervention is based on prior available, context-specific knowledge about possible effectiveness and the necessary preconditions (evidence-based).

In case the ToR identifies ‘impact’ as an evaluation criterion, the evaluation should also assess *unintended effects.* This includes the effects for the non-target population and the broader effects of the intervention on society, instead of only assessing the stated project objectives.

### **🡸 Click here to view the assessment grid criterion 7**

|  |  |  |
| --- | --- | --- |
| 7 | The ToR clearly identifies the relevant OECD/DAC evaluation criteria and cross-cutting topics that will be utilised to formulate the evaluation questions | **√√√ Good** |
| The ToR does not clearly identify the relevant OECD/DAC evaluation criteria and cross-cutting topics that will be utilised to formulate the evaluation questions | **ꓫ Inadequate** |

### **🡸 Click here to read an elaboration and examples of the OECD/DAC criterion ‘efficiency’**

|  |
| --- |
| The criterion ‘efficiency’ is often only mentioned or vaguely described in the ToR. In case the ToR mentions efficiency as an evaluation criterion, it should specify which aspects of efficiency will be considered in the evaluation.  The OECD-DAC evaluation ‘Efficiency’ distinguishes between different aspects: *economic* delivery of results, and *timely* delivery of results, including *organizational efficiency.*  **Economic efficiency**  When assessing economic efficiency, the evaluator can compare inputs (costs, staff time, natural resources) with outputs (efficiency in the strict sense), or with outcomes and impact (cost-effectiveness).  In some cases, outcomes and impact can be quantified in monetary terms, enabling a cost-benefit analysis. In cases where results cannot be quantified in monetary terms, it may be useful to compare the intervention with alternative interventions in terms of inputs needed to achieve similar results. This comparison is possible without quantifying the results.  Examples of good evaluation questions about efficiency:   * *‘How do intervention X’s quantified results at the outcome or impact level compare with the intervention’s costs?’*    + Cost-benefit analysis: benefits are quantifiable.   + This analysis can also be complemented with a comparison to alternative interventions * *‘How do intervention X’s (non-quantified) outputs, outcomes, or impact compare to costs, and how does this compare with alternative interventions?’*   + Efficiency or cost-effectiveness; benefits don’t need to be quantifiable   + Efficiency or cost-effectiveness is compared to that of alternative interventions   Examples of inadequate evaluation questions about efficiency:   * *‘To what extent is intervention X efficient?’*   + There is no specification of efficiency * *‘How do intervention X’s costs compare with its benefits?’*    + The benefits cannot be quantified;   There is no comparison with efficiency or cost-effectiveness of alternative interventions.  **Timeliness of results and operational efficiency:**  When assessing timeliness, the evaluator should assess to what extent the results have been achieved vis-à-vis the planned results, and how the actual expenditure compares to the planned expenditure, within the intended timeframe. The evaluation may also assess if efforts have been made to overcome obstacles and mitigate delays in implementation.  Evaluating timely delivery of results may also include an assessment of operational efficiency, which deals with the management of the intervention and the extent to which resources were well used during implementation. |

## **Evaluation questions**

The ToR should include a main research question and a number of evaluation questions in a precise and testable manner, aligning with the evaluation objective (criterion 5), the scope of the evaluation (criterion 6), and the relevant OECD/DAC evaluation criteria and cross-cutting topics (criterion 7).

Formulating overly broad or vague evaluation questions or mechanically translating the OECD/DAC criteria into questions should be avoided. At the same time, one should also steer clear of crafting too many detailed evaluation questions, which could dilute the focus of the evaluation. The evaluation questions should strike a balance between being ‘realistically ambitious’ while also considering the practical limitations of the evaluation, including constraints related to time, travel possibilities, budget, and the availability of information.[[4]](#footnote-5)

### **🡸 Click here to view the assessment grid criterion 8**

|  |  |  |
| --- | --- | --- |
| 8.a | The ToR formulates evaluation questions with a clear focus, systematically organized, logically stemming from the objective, scope, and the selected OECD/DAC criteria and cross-cutting topics | **√√√ Good** |
| The evaluation questions demonstrate some degree of focus and largely align with the evaluation’s objective, scope, and the selected OECD/DAC criteria and cross-cutting topics. | **√√ Adequate** |
| The evaluation questions lack a distinct focus and fail to logically align with the evaluation's objective, scope, and the selected OECD/DAC criteria and cross-cutting topics; | **ꓫ Inadequate** |
| The combined answers to all evaluation questions would not provide sufficient information to answer the main research question | **ꓫ Inadequate** |
| 8.b | There is minimal to no overlap in the evaluation questions and there are not too many detailed evaluation questions; | **√√√ Good** |
| The research questions exhibit some degree of overlap, but there are not too many detailed evaluation questions; | **√√ Adequate** |
| There are too many detailed evaluation questions; | **ꓫ Inadequate** |
| 8.c | Considering constraints such as time, travel possibilities, budget, and availability of information, the evaluation questions are realistic | **√√√ Good** |
| The evaluation questions are not realistic considering constraints such as time, travel possibilities, budget, or availability of information | **ꓫ Inadequate** |

# **Phase II – Elaborated methodology**

## **Research design**

The evaluation’s research design structures the overall approach and methods. The approach and methods must be appropriate answer the evaluation questions and achieve the evaluation’s objective in a valid and reliable manner. As the purpose of and possibilities for evaluation projects will differ, the research design may encompass multiple evaluation methods: how you evaluate depends on what you evaluate.[[5]](#footnote-6)

Evaluation robustness may be increased using multiple methods or data sources – a strategy known as ‘triangulation’.

|  |
| --- |
| **Triangulation** is a strategy to enhance the validity and reliability of the findings by cross-verifying information from different perspectives in order to provide a more comprehensive and accurate understanding of what is being evaluated. Triangulation can refer to the use of multiple:   * research methods; * data collection methods (i.e. interviews, surveys, observations); * data sources (i.e. persons, documents, project sites). |

The ‘inception report’ (or ‘technical report’) should elaborate on the research design, highlighting why the – combination of – method(s) have been selected and how they are expected to validly and reliably contribute to answering the research questions.

### **🡸 Click here to view the assessment grid criterion 9**

|  |  |  |
| --- | --- | --- |
| 9.a | The ‘inception report’ (or ‘technical report) articulates why the selected method(s) are appropriate to validly and reliably answer the various evaluation questions. | **√√√ Good** |
| The ‘inception report’ (or ‘technical report) mentions chosen research method(s), but the selected method(s) are inappropriate to validly and reliably answer the evaluation questions, or it remains unclear how the method(s) will contribute to answering them. | **ꓫ Inadequate** |
| The ‘inception report’ (or ‘technical report) does not elaborate on the research design:   * + - No or poorly described rationale for methods selected; or     - solely mentioning of data-gathering techniques; or     - solely mentioning information sources. | **ꓫ Inadequate** |
| 9.b | The research design includes a variety of research methods, data collection methods, and data sources, enabling effective triangulation of findings. | **√√√ Good** |
| The research design relies on a limited number of research methods, data collection methods, or sources, which hinders valid and reliable triangulation of findings. | **ꓫ Inadequate** |

## **10a. Plausibility of causal claim (qualitative)**

The selected evaluation method(s) should be appropriate to assess the contribution or attribution of the projects or interventions to observed results at the outcome or impact levels. Evaluators may use qualitative evaluation method(s) to evaluate the degree to which causal claims of projects or interventions about results, effects, and outcomes are plausible.

[White and Phillips (2012)](https://www.3ieimpact.org/sites/default/files/2019-01/working_paper_15.pdf) distinguish four qualitative evaluation methods that are suitable for substantiating claims about effectiveness (realistic evaluation; contribution analysis; process tracing and; general elimination methodology) and four methods that are less suitable for substantiating claims about effectiveness (most significant change; success case method; outcome mapping; method for impact). In recent years, outcome harvesting, which shares characteristics with outcome mapping, has also gained popularity amongst evaluators, especially for evaluating lobby and advocacy related activities.

Qualitative evaluation methods that are able to substantiate clausal claims about effectiveness should generally follow these five steps:[[6]](#footnote-7)

1. Formulate the cause-effect contribution question;
2. (Re)construct an intervention theory, including the assumptions;
3. Formulate alternative theories and explanations for the observed changes;
4. Collect data along results areas in the intervention theory, and for the alternative theories, including data from stakeholders that have not been directly involved in the project;
5. Verify in a step-by-step manner the causal chains of the intervention theory for the full range of possible outcomes (including achieved results, intended results that have not been achieved, and unintended effects), and the alternative theories.

Research designs may include participatory evaluation methods, which can be used to deepen the understanding of specific mechanisms and can improve the evaluator’s contextual understanding and to facilitate downward accountability.[[7]](#footnote-8) In order to answer evaluation questions about effectiveness it is important that, combined, the methods align with the five steps outlined above to validly and reliably answer the evaluation questions. Therefore, (single) participatory evaluation methods should be complemented with other evaluation methods. Furthermore, the methods should assess the full range of possible outcomes, including achieved results, intended results that have not been achieved, and unintended effects.

### **Click here for the assessment grid for criterion 10a**

|  |  |  |
| --- | --- | --- |
| 10-1 | The qualitative evaluation method(s) for effectiveness follow the five steps mentioned above. | **√√√ Good** |
| The method(s) follow steps 1 and 2, but do not include the identification of possible other factors affecting the results (step 3) prior to data-collection. The evaluator did consider other factors during and after data-collection during step 4; | **√√ Adequate** |
| The intervention theory, or ToC, is reconstructed only after the evaluator's data collection, with no distinct separation between formulating the cause-effect hypothesis and collecting data to verify or falsify it (steps 1 and 2) | **ꓫ Inadequate** |
| The evaluator did not consider alternative theories or explanations for the observed changes (step 3) | **ꓫ Inadequate** |
| There was no step-by-step validation along the causal chain, and the evaluator directly attributes results at the outcome level to project activities (step 5) | **ꓫ Inadequate** |

## **10b. Plausibility of causal claim (quantitative)**

Evaluators can use quantitative evaluation method(s) to robustly substantiate causal claims about effects of projects or interventions. The Maryland scientific methods scale delineates five progressing levels of rigour:

1. Single observation moment, after a project: a comparison with and without the project
2. Two observation moments, comparisons before and after the project, without a control group
3. Two observation moments: comparing before and after, and with and without the project (double difference).
4. Two observation moments: comparing before and after, and with and without the project (double difference, semi-experimental design), while accounting for other external influences.
5. Two observation moments: comparing before and after, and with and without the project (double difference); participants are randomly assigned to a project (randomized control group, experimental design).

Although Level 5 is best suited for attributing results to a specific project or intervention, that level of rigour is not always feasible for evaluations commissioned by the Netherlands Ministry of Foreign Affairs or its partner organisations. Alternatively, quantitative evaluations with a Level 4 rigour are generally accepted as robust to support substantial causal claims.

Evaluations below level 4 can be assessed as ‘adequate’, if the following thee assumptions are valid (see assessment grid):

1. Participants in the intervention group had similar values on the dependent variable (X) as those in the control group at baseline.
2. Participants in the control group exhibit similar (relevant) characteristics as those in the intervention group;
3. In the absence of the project, the dependent variable (X) would not change: there are no other factors that influence X.

### **🡸 Click here to view the assessment grid criterion 10b (quantitative)**

|  |  |  |
| --- | --- | --- |
| 10-2 | The method exhibits Level 4 or Level 5 rigour; | **√√√ Good** |
| The method exhibits Level 3 rigour, and the evaluator has actively validated the second assumption | **√√ Adequate** |
| The method exhibits Level 2 rigour, and the evaluator has actively validated the third assumption. | **√√ Adequate** |
| The method exhibits Level 1 rigour, and the evaluator has actively validated the first two assumptions | **√√ Adequate** |
| The quantitative evaluation method to evaluate effectiveness exhibits Level 3 rigour, but the evaluator has not actively validated the second assumption | **ꓫ Inadequate** |
| The method exhibits Level 2 rigour, but the evaluator has not actively validated the third assumption. | **ꓫ Inadequate** |
| The method exhibits Level 1 rigour, but the evaluator has not actively validated the first two assumptions | **ꓫ Inadequate** |

### **🡸 Click here to read examples for assessing criterion 10b (quantitative)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Examples ‘good’**   |  | | --- | | * 1. **Two observation moments: comparing before and after, and with and without the project; participants are randomly assigned to a project.** A cash transfer project intended, among other objectives, to enhance the nutritional status of participants. To assess the impact, households were randomly assigned either to receive cash transfers (intervention group) or to serve as a control group, not receiving the transfers. The evaluation entailed both before-after comparison and with-without comparison. Given the randomisation of participants, matching techniques were not necessary to safeguard that participants from the intervention and control groups display similar characteristics at the project’s outset.  1. **Two observation moments: comparing before and after, and with and without the project, while accounting for other external influences.** A rural electrification project intended, among other objectives, to increase the income of participants, generated at the household level. To assess this, the evaluators combined a before–after comparison (comparing baseline with endline) and a with–without comparison (intervention village versus control village). The evaluation deploys matching techniques to ensure that the intervention group and the control groups display similar relevant characteristics. |   **Examples ‘adequate’**   |  | | --- | | 1. **Two observation moments: comparing before and after, and with and without the project (double difference) with validated second assumption.** A rural electrification project aims, among other objectives, to reduce poverty levels among participating households. Participant selection is not based on household characteristics, but on geographical location, as the project is rolled out across districts over time. If the evaluator actively validates the assumption that participants in districts where the project has been implemented exhibit similar (relevant) characteristics to inhabitants of districts where the project has not yet been implemented, a double-different analysis can be considered adequate. 2. **Two observation moments, comparisons before and after the project, without a control group, with validated third assumption.** A rural electrification project aimed, among other objectives, to increase the number of hours children spend on doing homework. The hypothesis is that provision of electricity has resulted in children studying after dark. If the evaluator can actively validate the assumption that there were no other factors (e.g. money-earning activities) that affected the amount of time spent on homework, then not having a control group can be seen as adequate. 3. **Single observation moment, after a project: a comparison with and without the project, with three validated assumptions.** A rural electrification project aimed, among other objectives, to increase the amount of fresh food stored in household fridges. The evaluator must actively validate the following three assumptions:    1. There was no electricity before the project, both in the intervention group as in the control group – therefore the amount of fresh food stored was the same;    2. the two groups are comparable, e.g. in terms of the economic situation that might contribute to the amount of food people have stored.   In case the evaluator has actively validated the three assumptions, not including a baseline may be scored as adequate. |   **Examples ‘inadequate’**   |  | | --- | | 1. **Single observation moment, after a project: a comparison with and without the project, without three validated assumptions.** A project aims to enhance farmer productivity by providing farmers with modern agricultural machinery and equipment at reduced costs. If the evaluator measures if participating farmers have higher crop yields that non-participating farmers, without validating 2. If farmers in intervention and control area had the same level of productivity before the start of the project; 3. If farmers in intervention and control areas had similar (relevant) characteristics, such as motivation or risk-averseness;   Then the observed differences between participating and non-participating farmers cannot be attributed to the project.   1. **Two observation moments, comparisons before and after the project, without a control group, with invalidated third assumption.** A project aims to enhance farmer productivity by providing farmers with modern agricultural machinery and equipment at reduced costs. If the evaluator measures farmer productivity before and after the project, without a control group, and without validating if production would have been the same in the absence of the project, then the observed differences cannot be attributed to the project and the method has to be considered as inadequate. 2. **Two observation moments: comparing before and after, and with and without the project (double difference), with invalidated second assumption. A** project aims to enhance farmer productivity by providing farmers with modern agricultural machinery and equipment at reduced costs. If the evaluator measures farmer productivity before and after the project, for both an intervention and a control group, but does not validate if participating farmers have similar (relevant) characteristics that may affect their productivity, then the observed differences in a difference-in-difference analysis cannot be attributed to the project and the method has to be considered as inadequate. It might be the case that participating farmers are more motivated, entrepreneurial or less risk-averse than those not participating, because of the selection criteria, such as a required buy-in, applied by the project. | |

## **Indicators or description of result areas**

Formulating indicators or describing result areas at small intervals in the result chain can make an effect claim from activities to results more plausible. The ‘technical proposal’ or ‘inception report’ should propose valid quantitative indicators (see below), or qualitatively describe the result areas as concretely as possible to accurately measure what it claims to capture.

For quantitative evaluation methods, the indicators should be appropriate for assessing the intended results, and should adhere to the SMART criteria:

* **Specific:** indicators should be clear and specific, leaving no room for ambiguity. They should be linked to different levels in the ToC or results chain;
* **Measurable:** indicators should be quantifiable, allowing for measuring progress;
* **Attainable**: objectives and corresponding indicators should be attainable, making it possible to capture changes during the evaluation;
* **Relevant**: the indicators should align with the objectives of the evaluated project;
* **Time-bound**: there should be a defined timeframe, making it clear when the measurement takes place.

### **🡸 Click here to view the assessment grid criterion 11**

|  |  |  |
| --- | --- | --- |
| 11.a | The elaborated methodology proposes concrete indicators or describes results areas linked to all levels in the results chain. | **√√√ Good** |
| The indicators or results areas are linked to at least three levels in the results chain: intervention, intermediate result, and project result of interest; | **√√ Adequate** |
| The report proposes indicators for less than three levels in the results chain and, hence, reflect a too large step. | **ꓫ Inadequate** |
| 11.b | For quantitative evaluation methods, the report proposes indicators that adhere to the SMART criteria. | **√√√ Good** |
| For quantitative evaluation methods, the report proposes indicators that are ‘specific’ and ‘measurable’. | **√√ Adequate** |
| For quantitative evaluation methods, the report proposes indicators that are not ‘specific’, or not ‘measurable | **ꓫ Inadequate** |

## **Sampling and case selection**

Evaluations can include multiple sampling strategies and/or case selection strategies to serve different evaluation objectives or answer different evaluation questions.

**Sampling** is the process of selecting a subset of individuals or units from a larger population. The following distinction is important:

* Probability sampling (e.g. random sampling, stratified sampling, or systematic sampling), ensuring that each member of the population has a known chance of being selected. The objective is to create a representative subsample which allows the evaluators to generalise findings from the sample to the broader population and prevent selection bias – see box.

|  |
| --- |
| When the chosen sample in an evaluation does not accurately represent the entire target population, this is referred to as **selection bias**. It occurs, for example, when participants in an evaluation of a youth employment project self-select: more successful or more entrepreneurial young people may be more likely to voluntarily participate in an evaluation and, as a result, the findings might are not representative for all young people. |

* Non-probability sampling (e.g. convenience sampling, purposive sampling, snowball sampling) does not ensure that every member has a chance of being selected. Especially in qualitative research, evaluators apply non-probability sampling methods to study specific phenomena, without the objective to generalise results to the broader population.

In quantitative research, calculating the sample size should be done prior to data-collection and is ideally based on a power calculation.[[8]](#footnote-9)

**Case selection** choosing specific cases for in-depth study based on their relevance to the objective of the evaluation, and the evaluation questions. The main goal of case selection is to gain a deep, contextual understanding of (a) particular case(s), which may provide valuable insights. Depending on the objectives, evaluation questions and analysis strategy, evaluators may select typical cases, extreme cases, critical cases, or comparative cases.

When sampling and selecting cases, evaluators should:

* Formulate sample or case selection criteria (the set of characteristics that must be present) independently from the actively involved stakeholders;
* Be transparent about sample selection criteria applied, e.g. by presenting a list of all potential cases, interventions or countries and their scores or the selection criteria;

### **🡸 Click here to view the assessment grid criterion 12**

|  |  |  |
| --- | --- | --- |
| 12.a | The report provides a valid justification for all sampling and case selection strategies | **√√√ Good** |
| The report does not provide a valid justification for the combined sampling or case selection strategies | **ꓫ Inadequate** |
| 12.b | The evaluator formulated the sampling or selection criteria independently from the actively involved stakeholders and the report is transparent about the sampling criteria applied | **√√√ Good** |
| The evaluator did not formulate the sampling or selection criteria independently from the actively involved stakeholders, or the report is not transparent about the sampling criteria applied | **ꓫ Inadequate** |
| 12.c | Quantitative methods: sample size is based on a power calculation. | **√√√ Good** |
| In case power calculations are not feasible, the evaluator refers to the literature to identify the sample sizes used in comparable evaluations. | **√√ Adequate** |
| There is no consideration of the required sample size prior to data collection. | **ꓫ Inadequate** |

## **Sufficient and external information sources**

The ‘technical proposal’ or ‘inception report’ should present which information sources will be included in a way that minimizes source bias, by using sufficient and independent information sources.

|  |
| --- |
| **Source bias** occurs when the chosen information sources in an evaluation are not well balanced with certain selected sources emphasized over others. This can lead to unrepresentative selection of sources and skewed findings, affecting an evaluation’s validity. |

Collecting information from stakeholders that were not directly involved with the project can provide important information about external factors, assumptions or about what happens in the absence of an intervention. In addition to information sources obtained from direct stakeholders, such project staff or the targeted population,[[9]](#footnote-10) the evaluator should also include other information sources, such as:

* the non-targeted population;
* informed but not directly engaged stakeholders;
* the evaluator’s own direct observations;
* validated secondary data sources.

Because it may be difficult to predict the amount of data needed to reach the point of saturation, it is important that the evaluators have sufficient flexibility (time and resources) to add more information sources if needed.

### **🡸 Click here to view the assessment grid criterion 13**

|  |  |  |
| --- | --- | --- |
| 13.a | The proposed research design outlines a mix of information, obtained from directly involved and other information sources | **√√√ Good** |
| The report proposes to, largely or entirely, use information obtained from stakeholders directly involved with the project | **ꓫ Inadequate** |
| 13.b | There is enough flexibility for the evaluator to add more information sources, in case necessary to reach the point of saturation | **√√√ Good** |
| There is some flexibility (e.g. ‘unforeseen’ time, but no budget) to add additional information sources | **√√ Adequate** |
| There is no flexibility for the evaluator to add required information sources | **ꓫ Inadequate** |

## **14. Description of limitations and bias**

The 'inception report' or 'technical proposal' should clearly describe the limitations in the reliability and validity (both internal and external) of the proposed methodology, as well as any potential biases, including selection bias. Note that this criterion does not assess the limitations or bias itself, but rather the acknowledgement thereof. Criterion 17 assesses how the implications of bias are taken into account in formulating conclusions, and presents various examples.

|  |
| --- |
| **Reliability** is the degree to which an evaluation would consistently produce the same results under similar conditions. It involves the extent to which the evaluation, including data collection and analysis, can be replicated or repeated to yield similar results.  **Internal validity** refers to the trustworthiness and credibility of the findings within the specific context. It involves the extent to which the evaluation accurately describes relationships between the results and the intervention, minimizing the effects of other relevant factors. Various forms of *respondent bias*, for example, can affect the internal validity of an evaluation.[[10]](#footnote-11)  **External validity**, also known as generalizability, is the extent to which the results of a study can be generalized or applied to settings, populations, and conditions beyond the specific evaluation. *Selection bias*, for example, may affect the external validity of an evaluation – see criterion 12. |

### **🡸 Click here to view the assessment grid criterion 14**

|  |  |  |
| --- | --- | --- |
| 14 | The ‘inception report’ or ‘technical proposal’ transparently describes all potential limitations and bias that might arise from the proposed methodology, sampling strategy, case selection and available information sources. | **√√√ Good** |
| There is no reflection about limitations or bias | **ꓫ Inadequate** |
| There are obvious limitations or biases that are not discussed. | **ꓫ Inadequate** |

# **Phase III – Draft and final report**

## **15. Transparency**

Describing the analysis in a transparent manner allows for scrutiny, replicability, and a good understanding of the entire research process. The evaluation report should therefore present all research methods, data collection methods, data sources, and data analysis techniques employed in a systematic, transparent and complete yet accessible manner, either in the main report or in appendices to the report.

In case different research methods, data collection methods, or data sources provide contradictory findings, the evaluator should be transparent about the differences and the assessment thereof. Transparently presenting the range of evaluation findings prevents cherry-picking of certain results or the selective presentation of findings to fit a specific narrative. The report should explain how the evaluator has weighted and combined findings from sources and methods to come to an overall judgement and conclusion.

### **🡸 Click here to view the assessment grid for criterion 15**

|  |  |  |
| --- | --- | --- |
| 15.a | The evaluation report presents all research methods, data collection methods, data sources and data analysis techniques in a systematic, complete and transparent manner (e.g. in an evaluation matrix). | **√√√ Good** |
| The evaluation report mentions the most important research methods, data collection methods, data sources and data analysis techniques. | **√√ Adequate** |
| It remains unclear which research methods, data collection methods, data sources have been used. | **ꓫ Inadequate** |
| 15.b | The report (or annex) discusses and compares all findings from all different research methods, data collection methods and data sources. It transparently describes how it weighed the evidence and how it addressed discrepancies between findings. | **√√√ Good** |
| The report (or annex) presents the different results from the different research methods, data collection methods and data sources. The report describes how evidence has been weighed in general terms. | **√√ Adequate** |
| The report is not transparent about which results are from which research methods, data collection methods or data sources. It remains unclear how evidence has been weighed. | **ꓫ Inadequate** |

## **16. Answers to evaluation questions**

The draft and final report should provide an answer to all evaluation questions. Although the main and sub-conclusions may not be structured according to the evaluation questions, in principle they all must be answered in the text in an recognizable way. If evaluators faced unforeseen limitations during the evaluation process that prevented them to answer all evaluation questions, as a minimum the report must highlight which evaluation questions it did not answer and explain the reasons for this.

### **🡸 Click here to view the assessment grid for criterion 16**

|  |  |  |
| --- | --- | --- |
| 16 | The draft of final report provides an answer to all evaluation questions. | **√√√ Good** |
| The report does not provide an answer to all evaluation questions, but mentions which questions could not be answered and why. | **√√ Adequate** |
| The report does not provide an answer to all evaluation questions and does not mention which questions could be answered or why | **ꓫ Inadequate** |

## **17. Logic of conclusions and recommendations**

There should be a clear and rational connection in the evaluation report from findings to conclusions, and from the conclusions to the recommendations.

Each conclusion should be supported by specific findings. It is important that the final report takes possible limitations and bias (criterion 14) sufficiently into account. For example, the authors state that evaluation has will have limited external validity, the report’s conclusions should not generalized beyond the selected cases.

Recommendations, in turn, must logically follow from the conclusions presented in the evaluation report. If the ToR identified an action objective (criterion 5), the recommendations should be in line with the stated evaluation objective.

### **🡸 Click here to view the assessment grid for criterion 17**

|  |  |  |
| --- | --- | --- |
| 17.a | The conclusions follow logically from the research findings | **√√√ Good** |
| The conclusions do not follow logically from the research findings | **ꓫ Inadequate** |
| 17.b | The implications of the mentioned limitations or biases have been sufficiently taken into account in the conclusions; | **√√√ Good** |
| Implications of limitations or biases have not been sufficiently taken into account in the conclusions. | **ꓫ Inadequate** |
| 17.c | The recommendations follow logically from the conclusions presented in the evaluation report, and are in line with the stated evaluation objective. | **√√√ Good** |
| The recommendations do not follow logically from the conclusions presented in the evaluation report, or are not in line with the stated evaluation objective. | **ꓫ Inadequate** |

### **🡸 Click here to view examples for assessing criterion 17**

|  |  |  |  |
| --- | --- | --- | --- |
| **Examples ‘good’ taking into account limitations and bias**   |  | | --- | | Selection bias: The evaluator for a women empowerment project conducted interviews and focus group discussions in the final year. Because the evaluator did not succeed in interviewing sufficient women that dropped out of the project, she/he largely interviewed actively participating women. Largely excluding women who dropped out of the project may introduce selection bias, as the sample does not represent the perspectives of all women who were initially part of the project. Women that dropped out might have been less satisfied with the project.  The evaluator acknowledges that this is a limitation that affects the generalizability of the findings to the broader population of women involved in the project. In the report, therefore, the evaluator noted that the conclusions were only valid for actively participating women. |      |  | | --- | | Source bias: The evaluator for a project aimed at strengthening civil society conducted interviews and focus group discussions with supported organisations. The initial conclusion of the evaluator was that the project was relatively effective in facilitating the organisations’ service provision to the community.  Acknowledging the potential for source bias by only interviewing the supported organisations, the evaluator also interviewed informed non-participating stakeholders in the community. Doing this, the evaluator realised that civil society organisations in the area were organised largely along ethnic lines and primarily served their own constituencies. As the project did not include organisations from all ethnic groups, it did not serve the entire population. After interviewing a wide range of community members, the evaluator also concluded that the project had been effective facilitating services for some community members, but at the same time also contributed to increased ethnic tension in the community. |  |  | | --- | | Anchoring bias: An evaluator uses a quick survey among participants of an entrepreneurship project, conducted by the implementing organisation itself. The survey started by explaining the project's objectives, activities and expected results and then asked the subsequent two questions:   1. To what extent do you agree with the statement: *‘Compared to three years ago, I am now selling more vegetables on the market?* 2. To what extent do you agree with the statement: *‘The project made a substantial contribution to my ability to sell vegetables?*   The anchoring bias here arose from introducing various reference points (the project, its expected results and observed changes in sales) before seeking the participants’ opinions. These reference points could subtly influence respondents’ judgements, possibly leading to an unconscious overestimation. The evaluator acknowledged that data collected in this survey may contain (anchoring) bias, leading to overestimated results and, as a result, does not use the findings to investigate effectiveness. The evaluator may still use other interesting data obtained from the survey, e.g. about the design of the intervention. | |

# Annex 1. Assessment grid – Terms of Reference

**Main assessment**

|  |  |  |  |
| --- | --- | --- | --- |
| 1.a | The ToR presents a reference group with two or more independent and external members | **√√√ Good** | |
| The ToR presents a reference group with one independent and external member | **√√ Adequate** | |
| The ToR presents a reference group without independent and external members | **ꓫ Inadequate** | |
| 1.b | The ToR requests that the reference group will be involved in different stages of the evaluation (TOR, inception report, draft final report) | **√√√ Good** | |
| The ToR does not propose that the reference group will be involved in the different stages of the evaluation (ToR, inception report, draft final report). | **ꓫ Inadequate** | |
| 2 | The ToR requests that the evaluators should not possess involvement in the design or implementation of the project | **√√√ Good** | |
| If the ToR does not mention that researchers should be independent | **ꓫ Inadequate** | |
| 3.a | The ToR presents baseline data about project result indicators | **√√√ Good** | |
| The ToR does not present baseline data but includes an explicit strategy to assess progress and to (re)construct the baseline situation. | **√√ Adequate** | |
| There are no baseline figures on project result indicators and there is no strategy to assess progress or to (re)construct a baseline situation. | **ꓫ Inadequate** | |
| 3.b | The ToR presents a clear context and problem analysis which forms the rationale for the project’s interventions. | **√√√ Good** | |
| There is no (or insufficient) problem and context analysis | **ꓫ Inadequate** | |
| 4.a | The ToR presents an intervention logic, theory or ToC and addresses intermediate steps between inputs, activity, output, outcome, and impact. | **√√√ Good** | |
| The description of the intervention distinguishes between activities, outputs and outcomes. The results chain should make sense and not omit important steps. | **√√ Adequate** | |
| There is no specified or reconstructed description of the intervention, or it directly links implemented activities to results at outcome or impact level without distinguishing intermediate results. | **ꓫ Inadequate** | |
| 4.b | The ToR considers external factors that may influence outcomes at various levels in the results chain | **√√√ Good** | |
| The ToR does not mention any external factors that can influence outcomes at various levels in the results chain | **ꓫ Inadequate** | |
| 4.c | The ToR explicitly assumptions underlying the presented relationships at different levels in the results chain | **√√√ Good** | |
| The specified or reconstructed ToC does not mention any assumptions | **ꓫ Inadequate** | |
| 5 | The ToR clearly defines the evaluation period, the geographical focus and the expenditure included in the evaluation | **√√√ Good** | |
| The ToR does not clearly define the evaluation period, the geographical focus or the expenditure included in the evaluation | **ꓫ Inadequate** | |
| 6 | The ToR articulate a distinction between the knowledge and action objectives in the evaluation. The objectives are stated, and it is evident who will use the evaluation results and for what purpose. | **√√√ Good** | |
| The ToR has a clear knowledge objective, but it remains unclear how the findings will be utilised. | **√√ Adequate** | |
| The ToR solely mentions an action objective or aims to showcase effectiveness rather than evaluating it. | **ꓫ Inadequate** | |
| 7 | The ToR clearly identifies the relevant OECD/DAC evaluation criteria and cross-cutting topics that will be utilised to formulate the evaluation questions | | **√√√ Good** |
| The ToR does not clearly identify the relevant OECD/DAC evaluation criteria and cross-cutting topics that will be utilised to formulate the evaluation questions | | **ꓫ Inadequate** |
| 8.a | The ToR formulates evaluation questions with a clear focus, systematically organized, logically stemming from the objective, scope, and the selected OECD/DAC criteria and cross-cutting topics | | **√√√ Good** |
| The evaluation questions demonstrate some degree of focus and largely align with the evaluation’s objective, scope, and the selected OECD/DAC criteria and cross-cutting topics. | | **√√ Adequate** |
| The evaluation questions lack a distinct focus and fail to logically align with the evaluation's objective, scope, and the selected OECD/DAC criteria and cross-cutting topics; | | **ꓫ Inadequate** |
| The combined answers to all evaluation questions would not provide sufficient information to answer the main research question | | **ꓫ Inadequate** |
| 8.b | There is minimal to no overlap in the evaluation questions and there are not too many detailed evaluation questions; | | **√√√ Good** |
| The research questions exhibit some degree of overlap, but there are not too many detailed evaluation questions; | | **√√ Adequate** |
| There are too many detailed evaluation questions; | | **ꓫ Inadequate** |
| 8.c | Considering constraints such as time, travel possibilities, budget, and availability of information, the evaluation questions are realistic | | **√√√ Good** |
| The evaluation questions are not realistic considering constraints such as time, travel possibilities, budget, or availability of information | | **ꓫ Inadequate** |

# Annex 2. Assessment grid – Elaborated methodology

**Main assessment**

|  |  |  |  |
| --- | --- | --- | --- |
| 9.a | | The ‘inception report’ (or ‘technical report) articulates why the selected method(s) are appropriate to validly and reliably answer the various evaluation questions. | **√√√ Good** |
| The ‘inception report’ (or ‘technical report) mentions chosen research method(s), but the selected method(s) are inappropriate to validly and reliably answer the evaluation questions, or it remains unclear how the method(s) will contribute to answering them. | **ꓫ Inadequate** |
| The ‘inception report’ (or ‘technical report) does not elaborate on the research design:   * + - No or poorly described rationale for methods selected; or     - solely mentioning of data-gathering techniques; or     - solely mentioning information sources. | **ꓫ Inadequate** |
| 9.b | | The research design includes a variety of research methods, data collection methods, and data sources, enabling effective triangulation of findings. | **√√√ Good** |
| The research design relies on a limited number of research methods, data collection methods, or sources, which hinders valid and reliable triangulation of findings. | **ꓫ Inadequate** |
| 10-1 | | The qualitative evaluation method(s) for effectiveness follow the five steps mentioned above. | **√√√ Good** |
| The method(s) follow steps 1 and 2, but do not include the identification of possible other factors affecting the results (step 3) prior to data-collection. The evaluator did consider other factors during and after data-collection during step 4; | **√√ Adequate** |
| The intervention theory, or ToC, is reconstructed only after the evaluator's data collection, with no distinct separation between formulating the cause-effect hypothesis and collecting data to verify or falsify it (steps 1 and 2) | **ꓫ Inadequate** |
| The evaluator did not consider alternative theories or explanations for the observed changes (step 3) | **ꓫ Inadequate** |
| There was no step-by-step validation along the causal chain, and the evaluator directly attributes results at the outcome level to project activities (step 5) | **ꓫ Inadequate** |
| 10-2 | | The method exhibits Level 4 or Level 5 rigour; | **√√√ Good** |
| The method exhibits Level 3 rigour, and the evaluator has actively validated the second assumption | **√√ Adequate** |
| The method exhibits Level 2 rigour, and the evaluator has actively validated the third assumption. | **√√ Adequate** |
| The method exhibits Level 1 rigour, and the evaluator has actively validated the first two assumptions | **√√ Adequate** |
| The quantitative evaluation method to evaluate effectiveness exhibits Level 3 rigour, but the evaluator has not actively validated the second assumption | **ꓫ Inadequate** |
| The method exhibits Level 2 rigour, but the evaluator has not actively validated the third assumption. | **ꓫ Inadequate** |
| The method exhibits Level 1 rigour, but the evaluator has not actively validated the first two assumptions | **ꓫ Inadequate** |
| 11.a | | The elaborated methodology proposes concrete indicators or describes results areas linked to all levels in the results chain. | **√√√ Good** |
| The indicators or results areas are linked to at least three levels in the results chain: intervention, intermediate result, and project result of interest; | **√√ Adequate** |
| The report proposes indicators for less than three levels in the results chain and, hence, reflect a too large step. | **ꓫ Inadequate** |
| 11.b | | For quantitative evaluation methods, the report proposes indicators that adhere to the SMART criteria. | **√√√ Good** |
| For quantitative evaluation methods, the report proposes indicators that are ‘specific’ and ‘measurable’. | **√√ Adequate** |
| For quantitative evaluation methods, the report proposes indicators that are not ‘specific’, or not ‘measurable | **ꓫ Inadequate** |
| 12.a | | The report provides a valid justification for all sampling and case selection strategies, and is transparent about criteria applied | **√√√ Good** |
| The report does not provide a valid justification for the combined sampling strategy, or is not transparent about criteria applied | **ꓫ Inadequate** |
| 12.b | | The evaluator formulated the sampling or selection criteria independently from the actively involved stakeholders and the report is transparent about the sampling criteria applied | **√√√ Good** |
| The evaluator did not formulate the sampling or selection criteria independently from the actively involved stakeholders, or the report is not transparent about the sampling criteria applied | **ꓫ Inadequate** |
| 12.c | | Quantitative methods: sample size is based on a power calculation. | **√√√ Good** |
| In case power calculations are not feasible, the evaluator refers to the literature to identify the sample sizes used in comparable evaluations. | **√√ Adequate** |
| There is no consideration of the required sample size prior to data collection. | **ꓫ Inadequate** |
| 13.a | | The proposed research design outlines a mix of information, obtained from directly involved and other information sources | **√√√ Good** |
| The report proposes to, largely or entirely, use information obtained from stakeholders directly involved with the project | **ꓫ Inadequate** |
| 13.b | | There is enough flexibility for the evaluator to add more information sources, in case necessary to reach the point of saturation | **√√√ Good** |
| There is some flexibility (e.g. ‘unforeseen’ time, but no budget) to add additional information sources | **√√ Adequate** |
| There is no flexibility for the evaluator to add required information sources | **ꓫ Inadequate** |
| 14 | The ‘inception report’ or ‘technical proposal’ transparently describes all potential limitations and bias that might arise from the proposed methodology, sampling strategy, case selection and available information sources. | | **√√√ Good** |
| There is no reflection about limitations or bias | | **ꓫ Inadequate** |
| There are obvious limitations or biases that are not discussed. | | **ꓫ Inadequate** |

**Re-assessment for criteria 1 – 8:**

This phase also includes a re-assessmentof the first subset of criteria, to make sure whether the elaborated methodology is in line with the ToR. For each of the 8 criteria from the first subset, the following assessment grid guides the re-assessment:

|  |  |
| --- | --- |
| This criterion has already been positively assessed for the ToR, and the previous assessment is still valid. | **√√√ Good** |
| Although this criterion has already been positively assessed for the ToR, there have been changes in the setup or implementation of the evaluation. The evaluation still aligns with the ‘adequate’ requirements for this criterion. | **√√ Adequate** |
| Although this criterion has already been positively assessed for the ToR, there have been changes in the setup or implementation of the evaluation. As a result, evaluation no longer aligns with the ‘adequate’ requirements for this criterion. | **ꓫ Inadequate** |

# Annex 3. Assessment grid – Draft and final report

**Main assessment**

|  |  |  |
| --- | --- | --- |
| 15.a | The evaluation report presents all research methods, data collection methods, data sources and data analysis techniques in a systematic, complete and transparent manner (e.g. in an evaluation matrix). | **√√√ Good** |
| The evaluation report mentions the most important research methods, data collection methods, data sources and data analysis techniques. | **√√ Adequate** |
| It remains unclear which research methods, data collection methods, data sources have been used. | **ꓫ Inadequate** |
| 15.b | The report (or annex) discusses and compares all findings from all different research methods, data collection methods and data sources. It transparently describes how it weighed the evidence and how it addressed discrepancies between findings. | **√√√ Good** |
| The report (or annex) presents the different results from the different research methods, data collection methods and data sources. The report describes how evidence has been weighed in general terms. | **√√ Adequate** |
| The report is not transparent about which results are from which research methods, data collection methods or data sources. It remains unclear how evidence has been weighed. | **ꓫ Inadequate** |
| 16 | The draft of final report provides an answer to all evaluation questions. | **√√√ Good** |
| The report does not provide an answer to all evaluation questions, but mentions which questions could not be answered and why. | **√√ Adequate** |
| The report does not provide an answer to all evaluation questions and does not mention which questions could be answered or why | **ꓫ Inadequate** |
| 17.a | The conclusions follow logically from the research findings | **√√√ Good** |
| The conclusions do not follow logically from the research findings | **ꓫ Inadequate** |
| 17.b | The implications of the mentioned limitations or biases have been sufficiently taken into account in the conclusions; | **√√√ Good** |
| Implications of limitations or biases have not been sufficiently taken into account in the conclusions. | **ꓫ Inadequate** |
| 17.c | The recommendations follow logically from the conclusions presented in the evaluation report, and are in line with the stated evaluation objective. | **√√√ Good** |
| The recommendations do not follow logically from the conclusions presented in the evaluation report, or are not in line with the stated evaluation objective. | **ꓫ Inadequate** |

**Re-assessment for criteria 9 – 14:**

This phase also includes a re-assessmentof the second subset of criteria, to make sure whether the draft or final report are in line with the elaborated methodological plans. This means that the research design (criterion 9), the plausibility of the causal claim (criterion 10), the utilised indicators (criterion 11), the sampling and case selection (criterion 12), the information sources (criterion 13) and the description of bias (criterion 14) will be re-assessed. The following assessment grid guides the assessment:

|  |  |
| --- | --- |
| This criterion has already been positively assessed for the elaborated methodology. The draft or final report is in line with what the intentions presented in the elaborated methodology. | **√√√ Good** |
| While the this criterion has been positively assessed for the elaborated methodology, the draft or final report differs slightly from the intentions presented in the elaborated methodology. However, it still aligns with the ‘adequate’ requirements. | **√√ Adequate** |
| Despite the positive assessment of this criterion in the elaborated methodology, the draft or final report significantly diverges from the intentions presented and no longer meets the minimum requirements. | **ꓫ Inadequate** |

**Re-assessment for criteria 1 – 5:**

This phase also includes a re-assessmentof the first five criteria. The composition and role of the reference group (criterion 1) and the independence of the actual evaluators (criterion 2) will be re-assessed. In addition, the final report should also present a description of the context (criterion 3) and a description of the intervention (criterion 4). The report, finally, should be focus on the evaluation period, geographical focus and specific policy domain as requested in the ToR (criterion 5). The following assessment grid guides the assessment:

|  |  |
| --- | --- |
| This criterion has already been positively assessed for the ToR. The draft or final report is in line with what was asked in the ToR. | **√√√ Good** |
| While the this criterion has been positively assessed for the ToR, the draft or final report differs slightly from what was asked in the ToR. However, it still aligns with the ‘adequate’ requirements. | **√√ Adequate** |
| Despite the positive assessment of this criterion in the ToR, the draft or final report significantly diverges from what was asked in the ToR and no longer meets the minimum requirements. | **ꓫ Inadequate** |

1. The remainder of this document refers to ‘projects’ but the criteria are equally applicable to larger ‘programmes’. [↑](#footnote-ref-2)
2. Or policy theory, in case the evaluation is at policy level. [↑](#footnote-ref-3)
3. The final report must include a validation of the ToC (see criterion 13). It is important that the evaluator refrains from using the same data for (re)constructing and validating the ToC. [↑](#footnote-ref-4)
4. The ToR could consider the existing literature in the process of formulating evaluation questions to facilitate focussing on untested assumptions or activities without an evidence base. [↑](#footnote-ref-5)
5. See the ‘[Toolbox beleidsevaluaties](https://evaluaties.rijksfinancien.nl/toolbox)’, which provides a good guide on for designing evaluation research (in Dutch). [↑](#footnote-ref-6)
6. These steps can also be applied to evaluations of projects that focus on lobby and advocacy, even though the activities and results may not be pre-identified. Especially then, a comprehensive reconstruction of the ToC is an important step and should include all intermediate steps between activities, outputs, and outcomes, as well as an acknowledgement of external factors, and key assumptions (see also criterion 4). [↑](#footnote-ref-7)
7. Participatory evaluation methods engage different stakeholders – including the intended target population – in the evaluation process, with the aim to improve the data and/or empower the included stakeholders. [↑](#footnote-ref-8)
8. A power calculation is a statistical procedure used to determine the minimum sample size needed to detect a meaningful effect with a given level of confidence. See: [Schmidt et al, 2018](https://www.sciencedirect.com/science/article/pii/S0022202X18322243) [↑](#footnote-ref-9)
9. Representative from implementing organisations can facilitate contact between evaluator and target population, but cannot be present during interviews of focus group discussions. [↑](#footnote-ref-10)
10. Courtesy bias refers to the situation where a respondent tells you what they think you want to hear. In structured surveys, courtesy bias can affect both people’s reported behaviour and self-reported outcomes. Anchoring bias refers to the cognitive bias where respondents rely too heavily on the initial information provided (the "anchor") when answering subsequent questions. [↑](#footnote-ref-11)