IOB Evaluation quality criteria 2020

Introduction

Since long IOB has used evaluation quality criteria to monitor and assess the quality of evaluations and of the ToRs on which they are based. This document presents the latest version of these criteria. The criteria are organised around different stages of the evaluation process. Each criterion is explained and briefly illustrated with an example. The intended audience of this document are:

- Evaluators that actually do evaluation research;
- M&E staff of partners that are responsible for commissioning, managing and assessing the quality of external evaluations;
- MFA Policymakers and NL Embassy staff, responsible for commissioning and managing evaluations and for assessing the quality of evaluation reports of projects and programmes.

Reading guide

The following section of this document presents a brief list of the 26 evaluation criteria that have been grouped according to the various stages of the evaluation process. While we are aware that criteria can be relevant in more than one stage, for the sake of readability we have grouped them as follows:

- Criteria 1-9: these criteria can help in formulating a Terms of Reference. These criteria refer to the quality control of the evaluation, a description and background of the project or programme, the objective and delimitation of the evaluation and the evaluation questions. In the process of formulating a ToR, it is also advised to formulate an initial methodological plan (criteria 10-17).
- Criteria 10-17: these criteria can be used for reviewing and commenting on an inception report. These criteria focus on the methodological quality. At this stage, the methodological plan of the evaluation can still be adjusted if necessary.
- Criteria 18-24: these criteria can be used to assess the quality of the draft report under the assumption that the ToR and inception report have already been assessed. The criteria focus on the quality of the used methodology and on the conclusions. At this stage it is often no longer possible to adjust the implementation of the evaluation, but it remains possible to either complement the analysis or to reformulate conclusions given certain methodological limitations.
- Criteria 25 and 26: these criteria can be used for assessing the final evaluation report, under the assumption that the draft report has already been assessed.
- Criteria 1-26: For assessing the overall quality of the final evaluation report and the evaluation process.

This document provides an explanation per evaluation criterion and elaborates on the specific conditions that should be met for a criterion in order to be appreciated as *good, sufficient* or *insufficient.* For illustrative purposes, the explanation of various criteria is accompanied by practical examples.

Many of the criteria are interrelated and sometimes one aspect of an evaluation may affect several criteria. For example, if only the perception of directly involved stakeholders is used as an information source, this will negatively affect the scores for criteria 14 and 17.

It is important to note that the criteria have been developed mainly for assessing individual project evaluations, not for broader policy evaluations.

> When assessing the overall quality of the final evaluation report and the evaluation process, IOB recommends that at least 23 of the 26 evaluation criteria are scored as 'sufficient' or 'good'. In addition, there are 13 knock-out criteria. This means that if an evaluation scores 'insufficient' on that criterion, the evaluation as a whole should be regarded as insufficient. The know-out criteria are: 2, 4, 5, 10, 11, 13, 14, 15, 17, 20, 21, 22 and 23.[1] The knock-out criteria are accompanied by an Asterisk (*) in this document.

The list of 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 *

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 an 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

Elaboration of the evaluation criteria

Quality control of the evaluation

1. A reference group oversees the evaluation.

This group is composed of the commissioner of the evaluation, members with both thematic and evaluation experience, including at least one independent member. The role of the reference group is to assure evaluation quality and independence. It advises the commissioner on the Terms of Reference, the selection of evaluators, the elaborated methodology (inception report), and the draft evaluation report. Comments and advice from the reference group should be seriously taken into account by the evaluation team.

Good	*This criterion can be scored as good when:* *a. There is a reference group with at least one independent and external member; AND* *b. the reference group is asked for advise during the different stages of the evaluation (ToR, inception report, draft final report).*
Insufficient	*This criterion must be scored as insufficient when:* *c. There is no reference group at all; OR* *d. there are no independent and external member(s) in reference group; OR* *e. the reference group is not asked for advise in the different stages (ToR, inception report, draft final report).*

2. Evaluators are independent *

The evaluators and affiliated organisations have not been involved in the design or implementation of the intervention (project, programme, policy) under evaluation, and have no interest in the outcome of the evaluation.

Good	*This criterion may be scored as sufficient when:* *a. None of the evaluators have been involved in the design or implementation of the intervention; AND* *b. None of the evaluators is affiliated with one of the organisations active in the consortium responsible for the design or implementation of the intervention, or has been affiliated with one of the organisations during the implementation of the intervention; AND* *c. None of the evaluators has in the past worked for the Ministry of Foreign Affairs and has been responsible for formulating policy or setting up the programme that has led to the intervention under evaluation; BUT* *d. Programme staff facilitates contacts between external facilitators and beneficiaries and active stakeholders. Programme staff may accompany the external evaluators during field visits. Staff can help in making the necessary introductions, but is not present during interviews and does not play an active role in sampling or case selection.*
Insufficient	*This criterion must be scored as insufficient when:* *a. At least one of the evaluators has been involved in the design or implementation of the intervention; OR* *b. at least one of the evaluators is affiliated with one of the organisations active in the consortium responsible for the design or implementation of the intervention, or has been affiliated with one of the organisations during the implementation of the intervention; OR* *c. at least one of the evaluators has in the past worked for the Ministry of Foreign Affairs and has been responsible for formulating policy or setting up the programme that has led to the intervention under evaluation; OR* *d. The report does not mention whether or not evaluators were independent.*

Description and background of the intervention

3. Description of the context of the intervention

This can include the national, sector, and political context, and explains the rationale of the intervention.

Good	*This criterion can be scored as good when:* *a. There is elaborate collection of baseline data on the project-result indicators; AND there is a clear context and problem analysis.*
Sufficient	*This criterion may be scored as sufficient when:* *a. There is no baseline data, but a very elaborate problem- and context analysis, with a rationale for the intervention; OR* *b. In the absence of baseline data, there is an explicit strategy to measure progress and to (re)construct the baseline situation.*
Insufficient	*This criterion must be scored as insufficient when:* *a. There are no baseline figures on project-result indicators; AND there is no problem analysis*

4. Description of the intervention *

Preferably in a theory of change (ToC), otherwise an intervention logic or result chain. The evaluator may need to reconstruct a ToC, using whatever is available in project documentation, but with a critical reflection from the evaluator's point of view.

Good	*This criterion can be scored as good when:* *a. The (reconstructed) ToC, intervention logic or result chain is meticulously presented and takes intermediate steps between activity-output-outcome-impact into account; AND* *b. There is a discussion of external important factors that can affect results at different levels in the result chain; AND* *c. All assumptions between the presented relations are mentioned and there is an elaboration on the potential effects of the assumptions on effects at different levels in the result-chain.*
Sufficient	*This criterion may be scored as sufficient when:* *a. The (reconstructed) ToC, intervention logic or result chain is presented step-by-step and, at least, distinguishes between activities, outputs and outcomes. The result chain should make sense and should not omit important steps or factors; AND* *b. The most important assumptions between the presented relations are mentioned (e.g. between output and outcome and between outcome and impact).* *Example: 'The project aims to contribute to functioning and inclusive societies in the South by investing in the capacity of civil society organisations. Important assumptions for contributing to functioning and inclusive societies in the South through strengthening civil society organisations are:* *1. Civil society organisations legitimately represent the interests of the population;* *2. The general population recognises CSOs as their legitimate representatives;* *3. Civil society play an important role in influencing policymaking processes;* *4. Donors and implementing partners can identify the civil society organisations that can play a key role in influencing policymakers;* *5. Etcetera.'*

Insufficient	*This criterion must be scored as insufficient when:* *a. The (reconstructed) ToC, intervention logic or result chain is absent, or directly links the implemented activities to the results at outcome or impact level, without distinguishing intermediate results and assumptions.* *Example: the presented result chain can be summarised as follows without discussing intermediate steps or assumptions: 'The project aims to contribute to functioning and inclusive societies in the South by strengthening the capacity of civil society organisations.'*

5. Validation of the assumptions underpinning the ToC *

The evaluator does not take the project ToC for granted, but validates the ToC assumptions, which may refer to cause effect relations within the result chains, to the context, or to broader world views on development. The evaluator makes use of broader literature (reviews) to reflect on the validity of the ToC, and adjusts or reconstructs the ToC if necessary.

Good	*This criterion can be scored as good when:* *a. ToC assumptions are tested with high quality literature (preferably systematic reviews) before the evaluation; AND* *b. A clear strategy to test the underlying assumptions afterwards; AND* *c. In case of final evaluation report: assumptions are tested afterwards with evaluation results.*
Sufficient	*This criterion may be scored as sufficient when:* *a. ToC assumptions are tested as part of the ToR assignment with existing evaluations or literature; AND* *b. A clear strategy to test the underlying assumptions afterwards; AND* *c. In case of final evaluation report: assumptions are tested afterwards with evaluation results.*
Insufficient	*This criterion must be scored as insufficient when:* *a. The project logic or ToC is taken for granted, without critical reflection.* *b. There are no assumptions mentioned at all; OR* *c. There is no strategy to test the assumptions afterwards; OR* *d.*

Objective and delimitation of the evaluation

6. Description of the objective of the evaluation

Clarify what the evaluation results will be used for. There may be several objectives and it helps to distinguish:

a. a knowledge objective (knowing what works, how it works); this can be translated into knowledge questions that will result in conclusions.
b. an action objective (recommending what to do); this can be translated into policy questions that will result in recommendations. These objectives will also determine when the evaluation results are needed, e.g. for a next programme phase or new policy.

Good	*This criterion can be scored as good when:* *a. The objective of the evaluation is clearly mentioned and it is clear who will use the evaluation results and for what purpose.* *Example:* *- Knowledge objective: The objective of this evaluation is to determine and explain the level of effectiveness of programme X after five years of implementation.* *- Action objective: Insight in the effectiveness of policy programme X and the underlying reasons for this level of effectiveness should enable policy makers to decide whether or not it should be continued, and if it is continued, what improvements can be made.*
Sufficient	*This criterion may be scored as sufficient when:* a. *The objective of the evaluation is mentioned but it remains unclear what the results of the evaluation will be used for (thus, there is no distinction between knowledge and action objectives)* *Example: The objective of this evaluation is to determine the effectiveness of policy programme X after five years of implementation.*
Insufficient	*This criterion must be scored as insufficient when:* *a. The purpose of the evaluation remains unclear; OR* *b. There is only an action objective mentioned; OR* *c. The implicit objective is only to demonstrate, rather than investigate, the effectiveness.* *Examples:* *- Unclear: The goal of this evaluation is to see the contributions of policy programme X in a changing environment. (Contribution to what? What changes in environment?)* *- Only action objective: The objective of this evaluation is to gain input for the decision on the continuation of policy programme X.*

7. Delimitation of the evaluation

Clarify what part of the intervention, expenditure, period, or even what part of the ToC, is of interest for this evaluation.

Good	*This criterion can be scored as good when:* *a. The evaluation period is clearly defined; AND* *b. The geographical focus of the evaluation is clear; AND* *c. In case of various parallel result chains and various result levels, it is clear which result chains are covered in the evaluation and up to what level (e.g. outcome or impact)*
Insufficient	*This criterion must be scored as insufficient when:* *a. The evaluation period is not clearly defined; OR* *b. Geographical focus of the evaluation remains unclear; OR* *c. In case of various parallel result chains and various result-levels, it remains unclear which result chains are covered in the evaluation and up to which level (e.g. outcome or impact);*

Evaluation questions

8. Choice of OECD-DAC evaluation topics to be covered [1]

Based on the evaluation objectives and limitations, it may turn out that not all evaluation topics (relevance, effectiveness, efficiency, impact sustainability and coherence) are needed. This in turn will be reflected in the evaluation questions. (See OECD DAC evaluation topics 2019)[2].

Good	*This criterion can be scored as good when:* *a. A reasoned choice leads to the selection of OECD/DAC evaluation topics that will be used in the evaluation questions; AND*
Insufficient	*This criterion must be scored as insufficient when:* *a. It is unclear which of the OECD/DAC evaluation topics are used in the evaluation and why; OR* *b. It is clear that the OECD/DAC evaluation topics have not been used in formulating the evaluation questions; OR* *c. The evaluation topics are mentioned but this does not result in an organised and structured set of evaluation questions. Often, the evaluation topics are mechanistically translated in a set of universal evaluation questions but are not operationalised for the specific intervention or for the objective of the evaluation.*

[1] To avoid confusion between terms, this document refers to the OECD/DAC evaluation criteria as 'evaluation topics'.

[2] Besides the OECD criteria, other cross cutting subjects to be considered in the evaluation can be mentioned here, such as gender, poverty reduction, inclusiveness or climate smartness

9. Clear set of evaluation questions

The evaluation questions follow logically from the intervention under evaluation, evaluation objective and delimitation, and chosen evaluation criteria. Evaluation questions should not be too general or vague, but also not be too many and too detailed, losing focus.

Good	*This criterion can be scored as good when:* *a. There is clear focus in the evaluation questions. There is an organised set of evaluation questions that follows logically from the evaluation objective, delimitation and the chosen OECD/DAC evaluation topics; AND* *b. There is little to no overlap in the questions and there are not too many evaluation questions;* *Example of a good research questions (OECD DAC topic relevance):* *'To what extent were the activities aligned with the needs of the beneficiaries during the implementation of the project? Has the project been able to respond to changing needs as a result of the covid-19 pandemic?'*
Sufficient	*This criterion may be scored as sufficient when:* *a. There is a certain focus in the evaluation questions. The questions largely follow logically from the evaluation objective, delimitation and evaluation criteria; AND* *b. There may be some overlap between evaluation questions; OR* *c. The evaluation questions go beyond the evaluation objective or the main question.* *Example of a good research questions (OECD DAC topic relevance):* *To what extent were the activities aligned with the needs of the beneficiaries?'*
Insufficient	*This criterion must be scored as insufficient when:* *a. There is no clear focus in the evaluation questions. The questions do not logically follow from the evaluation objective, delimitation and evaluation criteria; OR* *b. The combined answers of all evaluation questions do not provide sufficient information to answer the main evaluation question; OR* *c. The evaluation questions are not formulated in a testable manner.* *Examples of research questions (OECD DAC topic relevance) without focus:* *1. To what extent was the project relevant?*

Evaluation methodology

A note of caution: our objective is to assess the evaluation methodology *as it has been conducted*, not only as it has been intended. This means that for example good intentions in the methodology chapter or an inception report will have to be verified in the actual results and conclusions chapters.

10. The research design is clearly elaborated and shows how the research results will contribute to answering the evaluation questions.

The design may consist of several quantitative and / or qualitative methods. If more than one method is used, the quality assessment looks both at the individual methods and the combination of methods.

a. Quantitative methods include three main research designs: survey, time series and experiment / quasi-experiment (see explanation under 11).
b. Qualitative methods are usually mainly based on elements of the Case Study approach and the Grounded Theory approach. Methods include many research designs, some of which are more suitable for evaluating effectiveness, and less susceptible for bias, than others. (see further explanation under criterion 11).

Good	*This criterion can be scored as good when:* *a. The research design has been clearly elaborated by showing what methodology or combination of methodologies are used including a clear explanation for why they have been chosen. AND* *b. The chosen methodologies are appropriate for answering the research question, and it is clearly explained how they do so.* *Example:* *RQ: How effective has the microfinance programme been, and what explains the level of effectiveness?* *In order to research the effectiveness of a microfinance programme for increasing household income and identify underlying reasons for this level of effectiveness, this research employs a mixed methods design. To determine the influence of microfinance on household income an experimental design will be implemented. This design will consist of before and after measurements in a treatment and similar control-group in order to isolate the effect of the microfinance programme and look for other explanatory factors. To understand how the microfinance programme works, and uncover explanatory mechanisms for the link between microfinance and household income a case study on the programme will be conducted, looking at various aspects of the programme's ToC. Together both methods will provide a complete picture on the level of effectiveness and underlying explanations for this level of effectiveness.* *[Of course further explanation and elaboration of more detailed methodological choices are needed on both designs. These are also highlighted in the following criteria.]*
Sufficient	*This criterion may be scored as sufficient when:* *a. The research design mentions the methodologies that are used and provides a sufficient explanation; AND* *b. The chosen methodologies are appropriate for answering the research question, and a link can be made to how they contribute to answering the research questions.* *Example:*

	RQ: How effective has the microfinance programme been, and what explains the level of effectiveness? *In order to research the effectiveness of a microfinance programme for increasing household income and identify underlying reasons for this level of effectiveness, this research employs a mixed methods design. To determine the influence of microfinance on household income an experimental design will be implemented. In addition, a case study on the programme will be conducted, looking at various aspects of the programme's ToC.*
Insufficient	*This criterion must be scored as insufficient when:* *a. The research design is not clearly elaborated. Consider the following elements:* *- Methodologies are mentioned but not/poorly explained* *- Only data-gathering techniques are mentioned (i.e. interviews, questionnaires, observations)* *- Only information sources are mentioned (i.e. type of respondents, documents); AND/OR* *b. The chosen methodologies are inappropriate for answering the research question, and/or it is unclear how they contribute to answering the research questions.* *Example:* *RQ: How effective has the microfinance programme been, and what explains the level of effectiveness?* *Lacking methodological explanation:* *In this research we will distribute a questionnaire amongst users of the microfinance programme to determine the level of effectiveness of the programme. In addition we will conduct expert interviews and focus-group discussions on the programme.* *Examples of inappropriate design choices:* *- A single measurement survey to determine effectiveness* *- A quantitative approach to uncover the underlying mechanisms of how the programme works* *- A case study to determine the level of effectiveness*

For criteria 11-15, a distinction is made between qualitative and quantitative methods, acknowledging that an evaluation often uses several methods.

11. The methods are appropriate to evaluate effectiveness: attribution and / or contribution (if effectiveness is an evaluation criterion/question) * [3]

a. ***Qualitative methods*** can make a plausible claim about the effect that the project has contributed to. The qualitative evaluation methods that allow a plausible claim have the following steps in common: (i) formulate the cause-effect contribution question; (ii) reconstruct an intervention theory; (iii) formulate an alternative theory; (iv) collect data along intervention and alternative theory; (v) validate the theories step by step.

A good overview of qualitative evaluation methods is provided by White and Philips (2012). They made an inventory of eight evaluation methods and distinguished four that make a more plausible claim of effectiveness:

1. Realist Evaluation;
2. Contribution Analysis;
3. Process Tracing; and
4. General Elimination Methodology.

This paper formulates a general framework for qualitative evaluation, using the four aforementioned methods. At the same time, the paper identifies four qualitative evaluation methods that are less suitable qualitative evaluation methods for making claims of effectiveness:

5. Most Significant Change;
6. Success case method;
7. Outcome Mapping;
8. Method for Impact Assessment of programs and Projects.

More recently, Outcome Harvesting has gained popularity amongst practitioners and evaluators as a qualitative tool for monitoring and evaluation. IOB recommends explicitly against the use of Outcome Harvesting as an independent, external evaluation method. In practice, this method is not appropriate to evaluate effectiveness and unable to validly establish the contribution of interventions to observed outcomes. In addition, the method is not in the spirit of several other evaluation quality criteria, specifically regarding the independence of evaluators (criterion 2), sufficient independent sources (criterion 17), triangulation (criterion 18), and avoidance of bias (criterion 19).

Qualitative methods

Good	*This criterion can be scored as good when:* *a. The evaluator first formulates a causal chain hypothesis before collecting data; AND* *b. The evaluator identifies possible other factors affecting the results, before collecting data. AND* *c. The evaluator validates both the causal chain, step by step, and checks the effects of other factors.* *Example: A project supports women with access to drinking water.* *(a) First, the evaluator formulated a hypothetical causal chain: project activities will result in an installed pump and in an organised women's group (outputs). One of the outcomes is that women would save money to pay for spare parts to maintain the pump. Another outcome is that the president of the women's group could contact the local water authority, in case they need spare parts for pump maintenance. The impact would be a maintained and functional water pump.*

[3] This corresponds with internal validity: To what extent is there a causal relationship between, for example, outputs and outcomes? (Vaessen et al, 2010, pp11-12)

	(b) Besides, the evaluator also identified other factors to check: are spare parts in stock at the water authority? Is the ground water table still high enough? Are women able and willing to pay for the part, if many other people use the pump as well? *(c) The evaluator then followed a step-by-step validation, which showed that the pump had been installed, that women were indeed organised, and their president indeed contacted the authority when the pump broke down. However, the evaluator also found out that the local authority did not have spare parts available for this type of water pump. The groundwater table was OK, and the women group had organised a fee to be paid by all users of the pump.* *The evaluator was thus able to draw conclusions about the project logic and about other factors affecting the results.*
Sufficient	*This criterion may be scored as sufficient when:* *a. The evaluator first formulates a causal chain (or uses an existing causal chain) as a hypothesis, AND* *b. The evaluator then collects data to validate this causal chain. AND* *c. The evaluator has considers other factors that may have affected the results, at least afterwards.* *Example: A project supports women with access to drinking water.* *(a) First, the evaluator formulated a hypothetical causal chain: project activities will result in an installed pump and in an organised women's group (outputs). One of the outcomes is that women would save money to pay for spare parts to maintain the pump. Another outcome is that the president of the women's group could contact the local water authority, in case they need spare parts for pump maintenance. The impact would be a maintained and functional water pump.* *(b) The evaluator then followed a step-by-step validation of this causal chain, which showed that the pump had been installed, that women were indeed organised, and their president indeed contacted the authority when the pump broke down. However, the pump was not repaired.* *(c) The evaluator checked possible other factors, and found out from the local authority that they did not have spare parts available for this type of imported water pump. He also checked the groundwater table, which was OK.* *The evaluator thus tested the intervention logic while keeping an eye on possible other factors.*
Insufficient	*This criterion must be scored as insufficient when:* *a. The causal chain between the intervention and the results is created after the evaluator's fact-finding (no separation between hypothesis and testing); OR* *b. There is no consideration of other factors affecting the results; OR* *c. Results at outcome level are attributed directly to activities without a step by step validation along the causal chain.* *Example: The evaluator did not use an intervention logic as basis for his evaluation, but asked women and local authority to explain why the water pump was not maintained. Women blamed the unwillingness of the local water authority to supply the spare part. The local water authority had doubts whether the ground water table was OK.* *The evaluator did not follow a project logic and did not consider other factors that were not mentioned by stakeholders.*

b. ***Quantitative methods*** can make a firm claim on the effect that can be attributed to the project. The Maryland scientific method scale distinguishes 5 levels:
    1. One observation moment, after the project: comparison with-without project.
    2. Two observations moments: comparisons before-after project, without control group.
    3. Two observation moments: comparing before-after AND with-without project (double difference).
    4. Two observation moments: comparing before-after AND with-without project (double difference, semi experimental design), and correcting for other, external influences.
    5. Two observation moments: comparing before-after AND with-without project (double difference); the participants are at random assigned to a project: randomised control group, experimental design).

    Level 5 is best suited for attributing results to a project, but is rare and not always possible to apply in evaluations. Level 4 is a commonly used good quantitative method. Level 1 and 2 are generally not the preferred methods for making effect claims and evaluators should be encouraged to aim at least for level 3 and preferably for level 4. Under certain strict conditions, evaluations below level 4 can be seen as just good enough, although in practise this is rare. Whether level 1, 2 or 3 is sufficient depends on the evaluation subject and context, especially on whether the following assumptions hold true: (i) that without the project nothing would change over time, and (ii) that a control group is similar to the project group, before the start of the project.

Quantitative methods

Good	*This criterion can be scored as good when:* *a. Level 4. The design considers the change that may happen without the project, and differences between participants and non-participants.* *Example: A rural electrification project. The effect of the introduction of electricity on the amount of income generated at the household level. To investigate this, the evaluators combine a before-after comparison (baseline – endline) and a with-without comparison (intervention village/control village), assuring that these two groups are comparable (use covariates or matching techniques).* *Level 5. Experimental design (randomised control group): project participants are randomly assigned to a project intervention and to a control group. A before-after comparison and with-without project comparison are combined (difference in difference).*
Sufficient	*This criterion may be scored as sufficient when:* *a. Level 1 + two validated assumptions. One observation moment, comparing with and without project, assuming that the control group was similar when the project started, and that nothing would have changed without the project. The evaluator has to actively validate these assumptions and demonstrate the evidence; OR* *Example: A rural electrification project. The effect of the introduction of electricity on the amount of fresh food stored in the fridge at home. The evaluation must validate the following two assumptions: (1) there was no electricity before the project and that nobody else would have introduced electricity in the area and (2) the two groups are comparable, e.g. on the general economic situation that might contribute to the amount of food people have stored*

	If both assumptions are validated then it may be seen as sufficient not to include a baseline. b. *Level 2 + one validated assumption. The study design compares the project group before and after the project, without a control group, and assumes that without the project nothing would have happened. The evaluator has to actively validate the assumption and demonstrate the evidence; OR* *Example: A rural electrification project. The effect of electricity on the number of hours children do their homework. Theory is that children spent time at night studying as a result of electricity. If the evaluation can validate and demonstrate that there were no other factors that affected the amount of time spent on homework (e.g. time spent on business activity), then not having a control group may be seen as sufficient.* c. *Level 3 + one validated assumption. The study design compares a project group with a control group, before and after the project, assuming that these two groups are similar. The evaluator has to actively validate the assumption.* *Example: A rural electrification project. The selection of participants is not affected by characteristics of the participants, but e.g. by geographical location. This may happen when a project rolls out its activities over different districts over the years. In this case, it may be assumed that the project participants are not better or worse performing than the control group (participant that have not yet been included in the project).*
Insufficient	*This criterion must be scored as insufficient when:* a. *Level 1 + one invalidated assumption. One observation moment, comparing with and without project, assuming that the control group was similar when the project started, and that nothing would have changed without the project; OR* *Example: A subsidised farm equipment project. Farmers that participated in the project obtain higher crop yields than farmers that did not participate in the project. We do not know if these assumptions hold or not, because perhaps these participating farmers may have had higher crop yields already before the project, or perhaps the more motivated and active farmers participated in the project and they would have yielded higher crops also without the project;* b. *Level 2 + invalidated assumption. The study design compares the project group before and after the project, without a control group, and assumes that without the project nothing would have changed; OR* *Example: A subsidised farm equipment project. Farm production has increased by 30%, and the evaluator attributes this to the project by comparing baseline data at the start of the project with endline data 3 years later. The assumption that production could not have changed in 3 years without the project cannot be validated. In fact, upon inspection it may appear that production outside the project area has increased as well. Therefore, we cannot attribute the changes in the project group area to the intervention;*

	c. *Level 3 + invalidated assumption. The study design compares a project group with a control group, assuming that these two groups are similar.* *Example: A subsidised farm equipment project. Data has been gathered from farmers at base- and endline for both the intervention and the control groups. However, it may turn out that farmers participating in the project turn out to be younger and more entrepreneurial than the people who don't participate, partly because the project was selective: farmers had to pay some cash up front, and partly because risk-adverse farmers were not keen to join the project. The assumption that participants are similar to non-participants can therefore not be validated and the measured differences between the groups over time cannot be attributed to the project.*

12. The methods are appropriate to evaluate efficiency (if this is an evaluation criterion/question)

The evaluation needs to specify what aspect of efficiency is considered[4].

a. Quantitative methods: e.g. calculation of cost-effectiveness, timeliness of implementation, overhead costs, etc.
b. Qualitative methods: e.g. organisational efficiency, assessment of demonstration or leverage effects and scaling, etc.

Qualitative methods

Good	*This criterion can be scored as good when:* a. *The methodology (a) explains how efficiency is considered, which matches the evaluation questions, and (b) evaluates accordingly.* *Examples:* 1. *Organisational efficiency (a) defined as slow and viscous procedures. (b) the evaluation analysed the organisation's main processes, and discussed strengths and weaknesses, and options for improvements: shortening the command chain.* 2. *Efficiency in leveraging effects (a) defined as the adoption of lessons from a pilot project being taken up in the national curriculum for farmer extension services. (b) The evaluation found that certain positive pilot results were known by some, and unknown by others, and have not yet been adopted in the national agricultural extension curriculum, because the recommendations were to complex and too expensive.*
Sufficient	*This criterion may be scored as sufficient when:* a. *The methodology (a) explains how efficiency is considered, and explains that this does not meet the expectations of the evaluation question in the ToR, due to lack of data.* *Example: The ToR asked for efficiency of the organisation, but evaluators were not able to investigate the project internal processes. Therefore, the evaluators made an inventory of project staff perceptions of strength and weaknesses of the organisational set up.*
Insufficient	*This criterion must be scored as insufficient when (same as under quantitative methods):* a. *Although the ToR has a question on efficiency without further specification, the methodology has no specification of how efficiency is considered. The results section on efficiency summarises ad hoc observations that indicate organisational efficiencies or inefficiencies.*

[4] The OECD-DAC evaluation criteria for efficiency considers the aspects (i) cost-effectiveness and (ii) operational efficiency, but there are more aspects of efficiency.

Quantitative methods

Good	*This criterion can be scored as good when:* *a. The methodology (a) explains how efficiency is considered, which matches the evaluation questions, and (b) evaluates accordingly.* *Examples:* *1. Organisational efficiency (a) defined as timeliness (or delays) of project implementation. (b) The evaluation found that a 3-year project had planned to train 2000 farmers each year, but training only started in the third year with 500 farmers, and asked for an extension.* *2. Efficiency of output delivery (a) defined as the unit costs of a training (10 half-day training) per trainee. (b) A comparison with benchmarks from literature about similar training found that the project training was relatively efficient.* *3. cost-effectiveness (a) defined as the monetary value of outcomes, compared with project costs. (b) The project invested $100 per farmer on training and a quality control system, resulting in a net additional profit of $50 per farmer each year.*
Sufficient	*This criterion may be scored as sufficient when:* *a. The methodology (a) explains that it was not possible to evaluate efficiency how this was questioned in the ToR, due to lack of data, but proposes something as relevant as possible using the available data, and (b) evaluates accordingly.* *Example: The ToR asked for cost-effectiveness of a farmer training, but evaluators were not able to collect data on the monetary value of the results, because most results, increased farm income, are to be achieved in the coming years. Therefore, the evaluators focused on comparing the unit costs (per trained farmer) with benchmark costs for farmer training.*
Insufficient	*This criterion must be scored as insufficient when:* *a. Although the ToR has a question on efficiency without further specification, the methodology has no specification of how efficiency is considered. The results section on efficiency summarises ad hoc observations that indicate organisational efficiencies or inefficiencies.*

13. The indicators or result areas are appropriate to capture the planned results along the different levels in the ToC * [5]

a. Quantitative methods: indicators are defined at different levels (e.g. output, outcome, impact; context and other assumptions) in the ToC. Indicators should be SMART and valid to measure the planned results.
b. Qualitative methods: result areas and processes, including assumptions that are part of the ToC, are defined at and between different levels (e.g. output, outcome, impact; context and other assumptions) in the ToC, and are valid to assess the planned results. Generally, qualitative result areas are a more open descriptions than SMART indicators used in quantitative research. Then the evaluator should still make them more specific during the evaluation.

Qualitative methods

Good	*This criterion can be scored as good when:* *a. Result areas are well described by operationalising underlying aspects/dimensions, and reflect the different result levels in the ToC at programme level between interventions and final results; AND* *Example: The evaluator uses first of all a documented or reconstructed programme theory of change with the following result areas: (i) capacity building of women's organisations; (ii) improved communication between women's group and local authorities; (iii) increased response by local authorities to women's needs.* *b. On a case-by-case basis: result areas of the overarching ToC are further refined into measureable results of a local project specific ToC, partly during project implementation, partly even during the evaluation itself, and often in discussion between evaluator, beneficiaries, project staff and other stakeholders.* *Example: The evaluator discusses with local stakeholders their project specific ToC, and identifies the following result areas: women's group identified a priority: the local district water authority should maintain the water pump. The women's group is made aware of the district planning processes, and of the key people at district level to lobby. The women's group chairman submits a proposal in time to district authorities, and discusses this with relevant district staff before the district decisions are taken. District decisions mention the women's priorities. The pump is maintained.*
Sufficient	*This criterion may be scored as sufficient when:* *a. Result areas describe at least three levels in the ToC: (1) intervention, (2) intermediate result, and (3) project result of interest; AND* *Example: (1) increased capacity of women organisations (2) increased interaction between women organisations and local authorities, and (3) Authorities respond to the needs of women.* *b. The ToC and its result areas, adapted by the project or reconstructed by the evaluator, are recognised by the beneficiaries and reflect their specific experiences.* *Example: The project translated the programme ToC in a project specific ToC: (i) women's communication and negotiation skills; (ii) communication frequency with local authorities; (iii) the*

[5] This corresponds with construct validity: To what extent is the element that we have measured a good representation of the phenomenon we are interested in? (Vaessen et al, 2010, pp11-12)

	district increases its budget for water pump maintenance. The evaluator discusses this with women, who recognise their project.
Insufficient	*This criterion must be scored as insufficient when:* *a. The criterion should be scored as insufficient when the used result areas reflect a too large step in the ToC; AND* *Example: by increasing the (1) capacity of women organisations, the project aims to contribute to enhanced (2) responsiveness of authorities to respond to the needs of women.* *b. The evaluation report does not show evidence of any reflection on the project specific ToC by final beneficiaries local implementing agencies, or other local stakeholders. It is therefore not clear whether capacity building of a women's organization addressed the specific capacity needs that hindered their further objectives.* *Example: The ToC describes women's group internal organisation as the main constraint, and therefore as the main output of the project. The evaluator did not check whether women recognise this and whether other factors were perhaps more relevant in this case, such as knowledge about district planning, or external relations between the women's group and district staff.*

Quantitative methods

Good	*This criterion can be scored as good when:* *a. Indicators are linked to different (intermediate) levels in the ToC (or result chain) between intervention and results; AND* *Example: A SRHR advocacy programme (without service delivery) that intends to reach and organise youth and provide awareness campaigns.* *In the ToC one of the result chains is reflected by the indicators: number of youth reached through awareness programme; number of youth clubs formed; frequency of meeting of these youth groups; frequency of contact between members; number of SRHR information campaigns channelled through these youth clubs; SRHR knowledge of adolescents; attitudes towards contraceptives; contraceptive use amongst adolescents in the previous 12 months.* *b. Indicators are linked to other parts in the ToC, or to non-project factors, that are likely to affect the project results; AND* *Example: % of youth indicating they can discuss SRHR topics with parents; % of youth indicating SRHR information is provided through schools; number of youth-friendly health providers in the area; access to contraceptives.* *c. Indicators are SMART (the above examples are SMART);* • *Specific (see a and b)* • *Measureable: it can be measured an unambiguous way.* • *Attainable, it is expected to change over the course of the project intervention.* • *Relevant: they say something about the intended results* • *Time bound: It is clear when it should be measured; when results would be achieved.*

Sufficient	*This criterion may be scored as sufficient when:* *a. Indicators are linked to at least 3 levels: project intervention, intermediate result, and project result of main interest; AND* *Example: A SRHR advocacy programme (without service delivery) that intends to reach and organise youth and provide awareness campaigns.* *Indicators: (1) number of youth clubs formed; (2) attitudes towards contraceptives; (3) contraceptive use* *b. The most relevant other factors, outside scope of project, are included if they are likely to affect the project results; AND* *c. Indicators are SMART, at least specific and measureable.*
Insufficient	*This criterion must be scored as insufficient when:* *a. Indicators are linked to only 2 levels in the ToC (or result chain); OR* *Example: A SRHR advocacy programme (without service delivery) that intends to reach and organise youth and provide awareness campaigns.* *In the ToC one of the result chains is only reflected by the indicators: (1) number of youth clubs formed and; (2) contraceptive use* *b. No indicators are included of factors that are likely to affect project results; OR* *Example: there is no indicator for the access to contraceptives* *c. Indicators are not SMART (see below)* *Examples of non-SMART indicators: an inclusive and youth-friendly society, a more liberal culture, enhanced rule of law, a vibrant local economy, land degradation, etcetera* *Examples of indicators that are not appropriate to measure the intended result: Measuring the number of youth that attended a training to make statements about strengthening gender equality, measuring the land area under sustainable maintenance to draw conclusions on soil degradation.*

14. Justified choice of sample, cases and information sources (e.g. choice of countries, projects, organisations and persons) * [6]

The sampling strategy or case selection should minimise selection bias.

Selection bias: the selected sample in the study is not representative of the target population, but rather correlates (positively or negatively) with project effectiveness and thereby undermines the generalisability of the findings. Self-selection of participants, for example, can play a role here: successful and more entrepreneurial beneficiaries may be more inclined to voluntarily participate in an evaluation of a youth employment programme.

a. Quantitative methods: Well justified choice of sampling strategy (e.g. random, stratified), (type of respondents, external validity), sample size (power calculation, response rate), and discussion of the limitations.
b. Qualitative methods: Well justified choice of the selection of cases and / or qualitative sample (based on strategic, theoretical or practical considerations), number of cases (internal validity, saturation), and discussion of the limitations.

Qualitative methods

Good	*This criterion can be scored as good when:* *a. The evaluator sets criteria for whom to interview, and asks project staff for information about organisations and/or for lists of beneficiaries, after which the evaluator makes the selection; AND* *b. Self-selection of respondents is greatly reduced. Besides the actively involved stakeholders, the evaluator makes an planned additional effort to include others that are less actively participating, have stopped participating or never participated; AND* *c. The selected cases follow logically from theoretical and/or practical selection criteria, and are appropriate to answer the research question which a (e.g. contexts or types of organisations); AND* *d. There is a justification of the sampling and case study selection, an explanation of how bias was as much as possible reduced, and justification of about the limitations and possible bias in the evaluation report.*
Sufficient	*This criterion may be scored as sufficient when:* *a. Project staff and evaluator have exploratory discussions about the criteria for selecting organisations and persons to interview, but the final selection is made by the evaluator; AND* *b. Self-selection of respondents is reduced: Besides the actively involved stakeholders, the evaluator makes an effort to include others that are less actively participating in the project; AND* *c. The selected cases capture most of the variety of the programme and hence are sufficient to draw valid conclusions; AND* *d. There is a justification of the sampling and case study selection, and a discussion about the limitations and possible bias in the evaluation report.*
Insufficient	*This criterion must be scored as insufficient when:* *a. Project staff determines the information sources, potentially resulting in biased findings; OR* *Example 1: the project staff prepares a list of villages, organisations and people to interview for the evaluator.* *Example 2; the staff helps in selecting which specific results of the programme should be evaluated.*

[6] This corresponds with external validity: To what extent can we generalise findings to other contexts, people, or time periods? (Vaessen et al, 2010, pp11-12)

	Example 3: the project staff prepares a selection of success stories, that the evaluator can validate. *b. Self-selection of respondents lead to bias; OR* *Example: Consultant visited a community asking people who were available to participate in a focus group discussion on the spot, resulting in a focus group with only the older men who were not working in the field during the day.* *c. The selected cases do not represent the project as a whole; OR* *d. There is no justification of the qualitative sampling strategy and case study selection, and no discussion about the limitations and bias of the sampling and case study selection in the evaluation report.*

Quantitative methods

Good	*This criterion can be scored as good when:* *a. The evaluator sets the criteria for case study selection. Project staff provides information about possible cases, and the evaluator makes the final selection; AND* *b. Self-selection of respondents is highly reduced; AND* *Example: if the woman in the household was not available, the household was skipped and the enumerator moved on to a neighbouring household.* *c. The sample size is based on a power calculation. This requires (a) one important indicator of interest, (b) the variance of that indicator and (c) the minimum effect size that the evaluation should be able to find. This calculation gives the minimum sample (e.g. 600 households) needed to detect if the effect if it is there; AND* *d. There is a justification of the sampling strategy and case study selection in the evaluation report.*
Sufficient	*This criterion may be scored as sufficient when:* *a. The evaluator sets the criteria for sampling. Project staff provides information and there is an exploratory discussion about case selection. Final selection should be made by evaluator; AND* *b. Self-selection of respondents is reduced; AND* *Example: if the woman in the household was not available for the interview, a second attempt is made the following day. If the woman is again not available, the questions for the women are skipped, and not answered by the man.* *c. The sample size was copied from similar studies that had shown significant effects. AND* d. *There is a justification of the sampling and case study selection in the evaluation report.*
Insufficient	*This criterion must be scored as insufficient when:* *a. Project staff plays an active role in selecting the sample and case studies, potentially resulting in biased evaluation findings; OR* *Example: guided by project staff, only the successful project villages close to the capital city were visited for the survey.* *b. The self-selection of respondents leads to bias; OR*

	Example: A survey included questions that have to be answered women. If the woman in the household was not available for the interview, the male head of the household responded – while women would have responded differently. *c. There is no consideration of sample size prior to data collection or the sample size is too small to find the intended effects; OR* d. *There is no justification of the sampling strategy and case study selection in the evaluation report.*

15. The analyses are appropriate, given the chosen research design.* [7]

a. Quantitative methods: appropriate statistical analyses, given the research design, chosen indicators and sample size; appropriate comparisons, e.g.: difference in difference, analyses of variance, regressions analyses, matching techniques.
b. Qualitative methods: the data analyses methodology is clear, given the research design, and includes e.g. theory construction, coding, comparing cases.

Qualitative methods

Good	*This criterion can be scored as good when:* *a. There is a clear analysis strategy which has been fully elaborated and applied. It shows how raw data leads to conclusions.* *b. Results from different information sources are presented transparently as presented in the methodology (note that information should not be traceable to individuals); AND* *Example: Farmers and staff appreciated training differently: Farmers had the impression that government maintenance tasks were delegated to farmers, while government staff had the impression that farmers could easily apply acquired knowledge.* *c. Results are transparently presented along the cause-effect chain (or theory), considering other factors, as presented in the methodology, step-by-step in separate text sections, and there is a critical reflection on the assumed theory.* *Example: Training of farmers in irrigation maintenance last year was appreciated by farmers. However, field observations showed that tertiary irrigation is still not maintained. Apparently the training was not sufficient to motivate farmers for joint maintenance. Some farmers who suffer less from water shortage point at the lack of water coming from upstream villages. Other farmers, who suffer more from water shortage, point at the unwillingness of farmers for joint maintenance. Government extension staff, who had not visited the village for over a year, has not followed up to discuss this problem with farmers.*
Sufficient	*This criterion may be scored as sufficient when:* *a. There is an analysis strategy which shows how raw data is translated in conclusions, but it is not completely elaborated, or applied, or it is difficult to see the link between data conclusions* *b. Results from different sources are aggregated when they converge, and presented separately when they don't converge, and are presented as announced in the methodology; AND* *Example: All agreed that the irrigation channels were poorly maintained and that some farmers receive too little water.*

[7] This corresponds with data analysis validity: To what extent are data analyses methods applied correctly for drawing adequate conclusions? (Vaessen et al, 2010, pp11-12)

	Although some farmers point at too little water coming from upstream villages, other farmers point at the lack of willingness in their own village for joint maintenance. c. *Results are presented along the result chain, considering other factors, following the methodology, and there is a discussion of the assumed theory.* *Example: First discuss the training, then joint maintenance, then the result in water availability, and discuss that the training was not sufficient to assure joint maintenance.*
Insufficient	*This criterion must be scored as insufficient when:* a. *There is no analysis strategy, which shows how raw data is translated into conclusions;* b. *The results do not transparently present the different information sources as presented in the methodology; OR* *Example 1: although the methodology distinguishes different information sources, e.g. opinions of farmers, government staff, and field observations, the report sums up the results without links to the sources: 'the main constraints were lack of water coming in the irrigation scheme, lack of maintenance, and lack of advices and training'.* c. *The results do not transparently present the information along the cause-effect chain (or theory), as presented in the methodology , but jumps to conclusions. There is no critical reflection on the theory.* *Example 'in spite of the training provided last year, farmers still experience a shortage of water in their fields'.* d. *There is no analysis at all*

Quantitative methods

Good	*This criterion can be scored as good when:* a. *The presentation of results is as planned in the design, and the results confirm that the design is indeed appropriate; AND* *Example: The design announced a difference in difference analysis (with-without and before-after project comparison), and results are indeed presented that way.* b. *The statistical analysis is as planned in the design, and the results confirm that the analysis is indeed appropriate.* *Example: the statistical analysis shows whether there is an effect or not; in case no effect was found, there is a discussion about the sample size and the minimum effect size that could be detected (power calculation).*
Sufficient	*This criterion may be scored as sufficient when:* a. *The presentation of results (and effect claim) is not in line with the chosen evaluation design, but this is explained in the report; OR* b. *The statistical analysis draws correct conclusions (effect claim), but does not answer the initial research question.* *Example: 'Although the means show a difference, the difference is not significant: the sample size was not large enough to measure a significant effect, so no conclusions about effectiveness can be drawn'.*

Insufficient	*This criterion must be scored as insufficient when:* *a. The statistical analysis (and effect claim) is not in line with the chosen evaluation design; OR* *Example: Although a double difference design (before-after and with-without project comparison) was planned in the methodology (or inception report), only before-after or only with-without results are presented, and interpreted.* *b. The statistical analysis draws false conclusions (effect claim).* *Example 1: 'Households have increased their income by 20%' while the statistical analysis is not done or the effect is not statistically significant.* *Example 2: 'The project had no effect on income', but the sample was too small to find a significant effect size.*

16. Summary of the methodology in an evaluation matrix

This matrix shows how (i) evaluation questions are translated into (ii) sub-questions / indicators / result areas, and (iii) methodologies and (iv) information sources.

Good	*This criterion can be scored as good:* *a. There is an overview in a table, that shows for each evaluation questions (a) indicators / result areas, (b) anticipated evaluation methods, and (c) information sources. AND* *b. The proposed indicators / result areas and information sources are convincingly sufficient to be able to answer each evaluation question.*
Sufficient	*This criterion may be scored as sufficient when:* *a. There is an overview, either in text or table, that shows per evaluation question, what indicators or result areas will be used, and what information sources will be used (methodologies may not be presented in the overview). AND* *b. The proposed indicators / result areas and information sources are likely to be sufficient to be able to answer almost all evaluation questions.*
Insufficient	*This criterion must be scored as insufficient when:* *a. There is no overview, not in text or a table, that shows how evaluation questions are 'translated' into (a) indicators or result areas, (b) methodologies, and (c) information sources. AND* *b. It is unclear whether the methodology will be sufficient to answer the evaluation questions.*

17. Sufficient independent information sources *

Besides information sources among project implementers, direct beneficiaries and other local stakeholders, the evaluator should also independently select and consult sufficient independent sources, e.g. the opinion of other experts or non-beneficiaries that can critically reflect on the intervention, objective observations, or validated secondary data.

Good	*This criterion can be scored as good when:* *a. A mix of dependent and independent information sources is used, including external people that can have a critical reflection on the project; AND* *b. Subjective information on perceptions is complemented with objective information (secondary, own field observations, measurements). AND* *c. The evaluator has used the opportunity to add more information when possible and relevant, aiming for information saturation (snowballing, adding additional documents, interviews during the evaluation)* *d. There is enough flexibility for the evaluator to adjust and add information sources during the evaluation, to reach saturation (planned extra time and support).*
Sufficient	*This criterion may be scored as sufficient when:* *a. A mix of directly involved (dependent) and independent information sources is used; AND* *b. There is some flexibility to add information sources during the evaluation (ad hoc squeezed in, without extra time).*
Insufficient	*This criterion must be scored as insufficient when:* *a. The main source of information is monitoring data, collected by the implementing organisation: OR* *b. The evaluation consulted no or very few independent informants that were not actively involved in project implementation or have an interest in the evaluation results; OR* *c. Representatives of the implementing organisations were present during interviews or focus groups with beneficiaries; OR* *d. There is no flexibility for the evaluator to adjust and add information sources during the evaluation, to reach saturation (no time, no possibility to re-plan).*

18. Triangulation of results from different information sources

This includes a comparison and critical reflection by the evaluator of results from different sources and results from different research methodologies, data collection methods (i.e. interviews, surveys, observations) and data sources (i.e. persons, documents, sites).

Good	*This criterion can be scored as good when:* *a. Sufficient different methodologies, data collection tools, and information sources are used; AND* *Example: a quantitative baseline –endline survey is combined with focus group discussions with farmers; farmer perceptions are combined with objective field observations; the opinions of project stakeholders are combined with the opinions of external experts* *b. The primary evaluation results are compared with other evaluations and a literature review (if available: a systematic review); AND* *Example: the project recommended practices (conservation agriculture) did not result in improved farm production in the 3rd year of the project. Literature review confirms that this is what would reasonably be expected in the first few years. ;*

	c. *The differences between results from different evaluation methods, data collection methods, and information sources are discussed. Besides, the results from the evaluation and the results from other evaluations and literature are compared and discussed.* *Example: although farmers in focus group discussions were positive about conservation agriculture, field observations revealed that farmers only apply this on a very small test plot. The survey data show an unexpected yield decline, which is confirmed in literature that shows that the main benefits of conservation agriculture are not short term yield gains, but long term yield stability and reduced labour and input costs.*
Sufficient	*This criterion may be scored as sufficient when:* a. *At least different evaluation methods, data collection methods and information sources are used, where this is possible. As a minimum: subjective perceptions are complemented by objective data (secondary data, observations); and opinions of project stakeholders are complemented by opinions of independent persons (experts not directly involved in the project, or not participating farmers); AND* b. *At least some use is made of other evaluations or literature about similar interventions. As a minimum: a reflection on the results of earlier evaluations is included; AND* c. *Results from different evaluation methods, data collection methods and information sources are presented in a disaggregated way, followed by a discussion.*
Insufficient	*This criterion must be scored as insufficient when:* a. *Not enough different methodologies, data collection methods and information sources are used to validly answer the main evaluation questions; OR* *Example 1: only quantitative farm production survey data are used, while qualitative focus group discussions could have added valuable information (or the other way round).* *Example 2: only farmer perceptions are asked, while field observations could have added valuable information (or the other way round).* *Example 3: only project staff and direct beneficiaries are interviewed about the project effects on farm production, while external government staff, agricultural experts, and non-participating farmers could have added valuable information as well;* b. *No use is made of secondary data, evaluations of similar interventions or broader literature reviews; OR* *Example: The evaluation may have found based on quantitative study of crop yield that the project recommended 'conservation agriculture' practices have not resulted in the (expected) good yields. No literature on 'conservation agriculture' practices is used to review what can be expected of these practices in the conditions in which they were implemented.* c. *There is no comparison and discussion about differences between results from different sources, research methods or data collection methods.* *Example: Farmers in focus group discussions were positive about the yield increased. Secondary yield data collected by the district agricultural department shows that yields did not increase (no comparison or discussion)*

19. Discussion of bias

The evaluator provides a critical reflection of different forms of bias (at least selection bias, respondent bias and evaluator bias). Note that this criterion does not reflect the extent to which bias is effectively addressed (C14), nor the extent to which bias has been taken into account in the presenting the conclusions (C23).

Selection bias: non-representative choice of cases and respondents (see Criterion 14).

Respondent bias: the problem of 'courtesy bias', whereby a respondent tells you what they think you want to hear, is well established. In structured surveys, courtesy bias can affect both people's reported behaviour and self-reported outcomes. Courtesy bias has a clear relevance for qualitative interviews, for example when interviewing respondents about how influential a particular agency or programme has been in affecting a change. A related form of bias is that of 'social acceptability' or 'political correctness' bias, where people provide responses which reflect what they regards as being the socially acceptable thing to say. (White and Phillips, 2012).

Evaluator bias: it is commonly held that there may be biases pushing evaluators towards positive findings, the main one being 'contract renewal bias'. Another bias which may actually be even more influential is 'friendship bias'. If an evaluator has spent time with programme staff and has developed a good relationship with them, it becomes difficult to upset them with a critical report (White and Phillips, 2012).

Good	*This criterion may be scored as good when:* *a. The evaluator is elaborate in discussing the possible biases the process of data collection- and analysis;* *Example:* *The evaluator acknowledges and describes that he was only able to discuss with a small group of women still actively participating in the project, and was unable to contact women that had stopped participating. This may result in a positive bias. We don't know why other women have dropped out: were they not satisfied?*
Insufficient	*This criterion must be scored as insufficient when:* *a. There is no critical reflection of biases; OR* *b. There appear to be more biases in the process of data collection and -analysis than discussed by the evaluator.* *Example:* *The evaluation acknowledges that only a small number of active women, present in the village during the evaluation, were contacted. However, although the evaluation report does present a high dropout rate, it does not mention that not contacting these women is a potential source of bias.*

20. Systematic, complete and transparent description of the data collection and analysis * [8]

In principle, if another evaluator would apply the same methodology, this should result in the same findings and conclusions (replicability).

Good	*This criterion can be scored as good when:* *a. There is a systematic, complete and transparent description of the data collection and analysis. As a result, it is very likely that if a different evaluator would follow this methodology, the evaluation would find very similar results.* *Example: survey questionnaire is included as annex, criteria for the selection for respondents are presented, the analysis is described and can be recognised in the way results are presented.*
Sufficient	*This criterion may be scored as sufficient when:* *a. There is a systematic, complete and transparent description of the data collection and analysis;[9] The methodology describes criteria for selecting respondents, includes guidelines (topics and guiding questions) for interviews and group discussions, and questionnaires for surveys (if applicable). As a result, it is likely that if a different evaluator would follow the same methodology, the evaluation would find more or less similar results.*
Insufficient	*This criterion must be scored as insufficient when:* *a. There is no systematic, complete and transparent description of the data collection and analysis. As a result, it is likely that if another evaluator would use the described methodology, the evaluator would find different results;* *Example 1: for a survey, the subjects are mentioned, but the questionnaire is not included.* *Example 2: the criteria used to select respondents (for a survey or for interviews) are not clear.* *Example 3: while both questionnaires and subjects for a survey are described, it remains unclear which methods have been used to process the collected data to arrive at the conclusions.*

[8] This corresponds with reliability: the transparency and replicability of the evaluation process (Vaessen et al, 2010, pp11-12)

[9] It is possible that the detailed methodology is included in the final evaluation report, in an annex, or in a separate inception report.

21. Discussion of the limitations of the evaluation *

The evaluator is self-critical and discusses the limitations of the study, including reliability, internal and external validity, relative contribution of the intervention and other external factors to the observed changes. Note that this criterion does not value the limitations of an evaluation, but rather the acknowledgment, discussion and implications thereof. As long as the limitations of a study are clear the study may still provide valuable information.

Good	*This criterion can be scored as good when:* *a. There is an elaborate discussion about all of the limitations of the evaluation;.* *Example* *In the methodology section, two limitations are discussed:* *(1) Only part of the project activities, provision of school meals, could be visited. Another part, hygiene awareness at home, could not be included in the evaluation. The evaluation can therefore only draw conclusions for that part of the project.* *(2) On ambitious outcome, reduced child malnutrition, could not be established by the evaluation. The evaluation can therefore not draw conclusions on reduced malnutrition.*
Sufficient	*This criterion may be scored as sufficient when:* *a. The evaluation briefly mentions the limitations of the evaluation and takes them into account sufficiently when presenting the findings and conclusions. In case the external validity is not mentioned the criterion may still be scored as sufficient as long as the report does not generalise the findings beyond the studied cases or implies to do so.*
Insufficient	*This criterion must be scored as insufficient when:* *a. There is no discussion about the reliability, validity or external factors that may have contributed to the findings in the evaluation; OR* *b. The limitations of the evaluation are mentioned, but the implications of these limitations are not sufficiently taken into account.* *Example:* *From the methodology or results section, it becomes clear that only part of the project activities (school meals) was visited, and that one outcome (reduced child malnutrition) could not be established. However, this is not discussed as a limitation.*

Results and conclusions

22. Conclusions answer research questions *

Although conclusions may be organised or grouped differently than the original research questions, in principle all research questions are answered, or accompanied by an explanation why they could not be answered.

Good	*This criterion can be scored as good* *a. All research questions are answered. The conclusions may be organised differently than the original research questions.*
Sufficient	*This criterion may be scored as sufficient when:* *a. Not all research questions are answered but it is explicitly mentioned which research questions could not be answered and why (for example, practical limitations that could initially not have been foreseen).*
Insufficient	*This criterion must be scored as insufficient when:* *a. Not all research questions are answered and there is no explanation why they were not answered.*

23. Conclusions follow logically from the research findings *

a. Complete and transparent presentation of the results of each method, to avoid jumping to conclusions. Detailed results can be presented in an annex.
b. Discussion of the limitations and validity of the conclusions (in line with C21)

Good	*This criterion can be scored as good when:* *a. All evaluation results are transparently presented, either in the evaluation report or in an annex; AND* *b. There is a (brief) reflection on the limitations of the evaluation; AND* *c. The presented conclusions are consistent with the methodologies used and all research findings, and take methodological limitations into account;* *Example of a limitation that is discussed, and taken into account in the conclusions:* *For a large youth employment programme, young men and women received business training and an investment capital to start a business – the exact amount depended on the quality of the respective proposal.* *In the methodology section the limitations are discussed. The evaluation was unable to contact women that had not participated in the above mentioned youth employment programme. We don't know why they did not participate: were they not interested, or was project support not offered to them? Moreover, we could not compare women that received business development support to women that did not receive this support, which would helped in attributing effects to the programme.* *In the conclusions sections, these limitations are taken into account: Women that received business development support have successfully increased their market sales by the end of the project. However, it is not clear to what extent this can be attributed to the evaluated programme. Because it remains unclear whether the project was inclusive to all interested women, or that only privileged or already entrepreneurial women*

	were able to participate. Moreover, we don't know whether the support has had a positive spill-over effect on other women, or that it has crowded out other women selling produce on the same market.
Insufficient	*This criterion must be scored as insufficient when:* *a. The results of the different methods are not transparently presented, either in the main report or in an annex; OR* *b. Not all the results of the evaluation are mentioned or taken into account in the conclusions, but rather, certain results or findings are presented to fit the narrative of the evaluator; OR* *c. Methodological limitations are not taken into account presenting the conclusions (see also criterion 19 and 21); OR* *d. Presented conclusions cannot be derived from the methodologies used and the data collected in the evaluation.* *Example of a limitation that is mentioned, but not taken into account in the conclusions:* *For a large youth employment programme, young men and women received business trainings and an investment capital to start a business – the exact amount depended on the quality of the respective proposal. After finalisation of the programme, the evaluators decided to only perform several in-depth studies for some of the larger successful individual businesses.* *In the methodology section, the evaluators concede that they incorporate selection bias not only by focussing on success stories, but also on the larger projects. This pragmatic approach, they argue, can help them in formulating certain conditions for success, and thus facilitates learning.* *When discussing the findings and conclusions of the study, however, the evaluators still made claims about the effectiveness of the entire programme, while the selection bias prevents them from generalizing their findings for the entire group of beneficiaries.*

24. Validation of draft conclusions

To strengthen the validity of the conclusions, the draft conclusions are discussed, e.g. in a validation workshop, with project implementers, independent experts, and compared with findings in earlier evaluations and broader literature.

Good	*This criterion can be scored as good when:* a. *In addition to the reference group meetings, the evaluators validate the findings with independent experts; AND* b. *The draft conclusions are discussed with the project implementers prior to publication; AND* c. *The conclusions are clearly positioned in the broader context of findings from previous evaluations of similar interventions and, if available, high-quality research such as impact evaluations and systematic reviews. In case the findings from the evaluation deviate from the existing knowledge base, the evaluators must reflect on the differences in findings and conclusions.*
Sufficient	*This criterion may be scored as sufficient when:* a. *The draft conclusions are discussed with the project implementers prior to publication; AND* b. *The conclusions are compared to findings from previous evaluations of similar interventions and, if available, with high-quality research such as impact evaluations and systematic reviews.*
Insufficient	*This criterion must be scored as insufficient when:* a. *The draft conclusions are not discussed with project implementers prior to publication. Note that the content or the interpretation of the findings are not up for discussion in such validation meetings, although factual inaccuracies may be mentioned by implementing actors. These meetings should be used to formulate useful and realistic recommendations, that are within the sphere of influence of the implementing agencies; OR* b. *The conclusions are not compared with previous evaluations of similar interventions or, if available, with high quality research such as impact evaluations or systematic reviews.*

Usefulness an readability of the evaluation report

25. Recommendations should be useful and practical, given the evaluation objectives and its intended users

Good	*This criterion can be scored as good when:* *a. The recommendations reflect the objective, either knowledge or action, of the evaluation; AND* *b. The recommendations follow logically from both the conclusions and the findings of the evaluation; AND* *In case of an action objective, the recommendations match the sphere of influence of the user(s) and the recommendations are realistic.*
Insufficient	*This criterion must be scored as insufficient when:* *a. The recommendations are not in line with the stated evaluation objective (see criterion 6); OR* *b. The recommendations do not follow logically from the findings and conclusions of the evaluation; OR* *c. In case of an action objective (criterion 6), there is no clear action perspective in the recommendations; the recommendations do not match the sphere of influence of primary user(s) of the evaluation; OR*

	Example: an evaluation in a fragile country recommends increasing focus on physical well-being, because the study finds that human security is an important prerequisite for strengthening the rule of law. The actual well-being of the population, however, is beyond the sphere of influence of the project implementers. *d. In case of an action objective (criterion 6), the recommendations are not realistic.* *Example: an evaluation finds that sexuality education can be an effective tool for influencing norms and values regarding sexual rights of minorities and recommends scaling up the activity without considering the increasingly restrictive political climate for such activities.*

26. The report is well readable, consistent, and includes a clear summary with evaluation objective, evaluation questions, conclusions and recommendations

Good	*This criterion can be scored as good when:* *a. The summary includes the (1) objective, (2) short policy description, (3) main evaluation questions, (4) main findings and conclusions that cover the main evaluation questions, (5) recommendations. OR* *b. In the main report, there is a consistency between evaluation questions --> methodology --> results --> conclusions, and the conclusions answer the evaluation questions.* *c. The text is well written and unambiguous.*
Insufficient	*This criterion must be scored as insufficient when:* *a. There is no summary. OR* *b. The summary does not include the (1) objective, (2) short policy description, (3) main evaluation questions, (4) main findings and conclusions that cover the main evaluation questions, (5) recommendations. OR* *c. In the main report, there is an inconsistency between evaluation questions and methodology, OR between methodology and results, OR between results and conclusions, OR between conclusions and evaluation questions.* *d. The text is not easy to read, and sometime ambiguous.*