Norad

EVALUATION DEPARTMENT

Report 3 / 2021

ANNEX 3–7

Annual Quality Assessment of Reviews in Norwegian Development Cooperation (2019–2020)

Annual Quality Assessment of Reviews in Norwegian Development Cooperation (2019-2020)

Commissioned by:

Norad Evaluation Department

Carried out by:

Ternstrom Consulting AB

Written by:

Ingela Ternström (team leader)

Jock Baker, Stefan Dahlgren, Eva Lithman and Abid Rehman (team members)

Abhijit Bhattacharjee (quality assurance)

October 2021

Ternstrom Consulting

Table of contents

List of Figures

List of Tables:

Acronyms and abbreviations

ACT	Agricultural Council of Tanzania
DAC	Development Assistance Committee (OECD)
DGRU	Democratic Governance and Rights Unit (University of Cape Town)
MFA	Ministry of Foreign Affairs (Norway)
Norad	Norwegian Agency for Development Cooperation
OECD	Organisation for Economic Co-operation and Development
PFS	Partnership for Scale Programme
TAP	Tanzanian Agricultural Partnership
TOR	Terms of Reference
UN	United Nations
UNEG	United Nations Evaluation Group

Annex 3: Methodology

Annex 3.1: Evaluation matrix

Table 1: Evaluation matrix

Evaluation objective	Data collection and analysis method
Assess the quality of reviews and decentralised evaluations of Norwegian development cooperation	Pilot test and calibrate the scoring tool Apply quality assessment tool (as per the rating manual) to each review and terms of reference; continued checks and support to ensure consistent scoring Develop database of scores and comments Analyse scores to identify areas of low and high quality; compare with terms of reference scores Present descriptive statistics on quality of reviews and discuss aspects of overall quality for different criteria
Identify strengths and weaknesses of reviews and decentralised evaluations	Analyse scores on each criterion to identify areas of high and low evaluation quality Analyse correlation to identify relationship between quality for different criteria and possible explanatory factors (e.g. types of reviews, thematic areas, commissioner); analyse justifying comments to identify causes of high and/or low quality and examples of good practice Translate results of analysis of scores into text describing strengths, weaknesses and potential explanatory factors (correlation)
Summarise findings from the reviews and decentralised evaluations, taking into consideration the credibility, based on assessed quality, of reviews	Discuss with EVAL if there are priority areas of interest (e.g. gender or strategy-related findings) Include tool for collecting data on main findings in scoring tool and database, collect findings data from all reviews (as part of review process), collate findings data in database Develop a "quality index" based on assessed quality of key criteria and decide index cut-off point (sufficient quality) for inclusion of findings in annual analysis Analyse and presentat findings for sufficient-quality reviews; compare findings for high- and low-quality reviews to assess if there is correlation between quality and findings

Annex 3.2 Ethical considerations, risks and limitations

Table 2: Ethical considerations and safeguards

Ethical consideration	Safeguards
Negative reactions to assigned scores may affect future employment opportunities for raters. This risk is increased by the decision to share scoring protocols with commissioners and evaluators.	Information that reveals who has quality assessed which evaluation will not be published. This information will only be shared with EVAL.
Misinterpretation of assigned scores, incorrect scores or poorly explained scores may negatively affect future employment opportunities for evaluators.	The sharing of scores and comments increases the importance of delineating justification and formulation in comments, and we will pay greater attention to this.

	To add reliability, we suggest, as an option, that a share of the reports be double scored by the team leader. Information about the quality of individual reports will only be shared together with justifying comments to explain the scores and will only be shared with the commissioner and evaluator of each report. All information will be shared with the Evaluation Department, which may share it with the quality assurance units at Norad and the MFA.
The scoring of evaluation report quality may emphasise the quantitative aspect of the quality assessment process, at the cost of learning and understanding of the reasons for the scores.	The 2021 annual quality assessment report will have a stronger focus on explaining the reasons for the assigned scores and provide examples of high and low quality to support learning. Evaluation commissioners and implementers will be encouraged to review the scoring protocols to better understand the quality that has been assessed.
The quality assessment process may make commissioners more reluctant to share reports and terms of references if they fear the quality is low. This may have a negative impact on the extent to which reports are made publicly available.	Ensure that all reports, published or not, are included in the quality assessment. This requires assistance from Norad and the MFA in requesting that reports and terms of references are shared with the quality assessment team.

Table 3: Risks and mitigation strategies

Risks	Consequences and/or mitigation strategies
Inter-rater bias: The risk that not all raters use the quality assessment tool in the same way.	Reduced by consecutive pilot scoring and team discussions of three evaluations. After the scoring of each pilot report the team will discuss scores and calibrate the use of the tool; double scoring of 5-10 percent of the reviews; conference calls as needed to discuss problems and solutions; comparison of scores and justifying comments and analysis of scoring data.
Intra-rater inconsistency: The risk that a single rater does not use the tool consistently across reviews.	Reduced by the short time period for scoring. It will also be recommended that raters do quality assessments in batches and start each batch by looking at some past quality assessments.
Evaluator bias: The risk that the identity of the report authors affects the score.	The risk will be discussed in the team meeting; To avoid conflicts of interest, distribute reviews by the same author to different team members; and cross-check outliers and suspicious trends during analysis of rating data.
Report structure bias: The risk that the structure and language of a report affect the assessment of the quality of the content.	Raters are made aware of this risk during the pilot scoring and will be reminded to focus on content and disregard, to the extent possible, how it is presented.
Attempts at influencing raters to assign high quality ratings: Various parties to the evaluation process may have an interest in affecting the quality rating.	Decreased by not publishing information that reveals who has quality assessed which reviews and by stating this explicitly in team meetings and reports. This information will only be shared with EVAL.
The number of evaluations constitutes an unknown share of all evaluations. This makes it difficult to draw conclusions about representativity.	We cannot assess the representativity of ratings or findings to other reviews. We will not be able to control for this. Findings and conclusions need to be interpreted with this limitation in mind.
The reports shared with the team may be of higher quality than reports that are unpublished or not shared.	A comparison of ratings across evaluations published on Evalueringsportalen or Norad.no and other evaluations may give an indication of “publishing bias”, but we cannot affect this. The number of evaluations published online is, however, too small for such a comparison.

Annex 3.3 Request for evaluation reports

Letter requesting review and evaluation reports

Dear All,

We have been tasked by the Evaluation Department in Norad to assess the quality of reviews and decentralised evaluations. Our work is part of an ongoing process to improve the quality of evaluations and reviews. This is important as the findings and recommendations of such evaluations are being used for decisions regarding programming and financial support and for reporting on results.

Thanks to your invaluable help we were able to access a large number of evaluation reports last year. Our report (Evaluation Department 06/2020) shows that although there have been some improvements over the past years, the quality of evaluations is still low.

We are now in the process of collecting evaluations for the next round of quality assessments, and again depend on your help in locating and accessing these. We realise that many planned evaluations may have been cancelled or postponed last year but still hope there are a few reports out there.

We are looking for reviews and evaluations that were:

- Finalised during 2020 or 2019 (unless included last year: a list of included reports is available in Annex 6 of the report)
- Mid- or end- term reviews or evaluations of projects, programmes, interventions etc. funded via Norwegian development cooperation
- Commissioned by Norad, MFA or Norwegian Embassies (including evaluations commissioned on behalf of the Ministry of Climate and Environment)
- Carried out by internal, external or mixed teams
- Carried out with or without field visits

Please make sure to include both main report, annexes and terms of references and send these to ingela@ternstromconsulting.se, no later than January 20th.

Please also let us know if your department/embassy does not commission evaluations or if planned evaluations were postponed or cancelled.

Please don't hesitate to email me if you have questions: ingela@ternstromconsulting.se.

Questions regarding the assignment can be directed to the Evaluation Department, attn; Ida.Lindkvist@norad.no and Tove.Sagmo@norad.no.

Thank you for your kind collaboration!

Ingela Ternström
Team Leader, Annual Assessment of the Quality of Reviews in Norwegian Development Cooperation

****

Ingela Ternström
Ph.D. Economics
Senior Consultant
Ternstrom Consulting AB
www.ternstromconsulting.se

Annex 3.4: Data collection tools

Revised scoring protocol format for reports

General info	Report Id (as in document name, e.g. 1805RT)		
	Name of assessor (initials, e.g. IT)		
	Date of assessment (yymmdd)		
	Assessment no. for the assessor (e.g. 1, 2, etc)		
	Time spent, *approx.* hours		
	Type of report (mid, end, etc.)		
	ToR was also reviewed (Y/N)		
Key quality Criteria	**Quality statements**	**Score**	**Comments explaining the choice of score**
Quality area 1: summary, style and structure			
1.1 Executive summary *Note: Be open-minded as to what you regard as a summary. If it is called something else but seems to be a summary, assess it.*	**The report contains a complete, accurate and concise executive summary. Examples of what a complete executive summary could include is rationale, purpose, scope, evaluation questions, brief description of methods and limitations, conclusions and recommendations.** Score 1: An executive summary is not provided. Score 2: The executive summary is inadequate with major gaps. Score 3: The executive summary is adequate, but with minor gaps. Score 4: The executive summary is complete, accurate and concise.		
1.2 Style and structure	**The report is clearly written and properly edited. Editing refers to both grammar and flow of information, where each section builds on the previous sections with no jumps or gaps in information.** Score 1: The report is not clearly written. Score 2: The report is partially accessible, but with poor editing. Score 3: The report is accessible, with minor editing gaps. Score 4: The report is clearly written and properly edited.		
Quality area 2: review purpose, objectives, and scope			
2.1 Purpose of the evaluation	**The rationale, purpose, intended users and intended use of the evaluation are stated clearly, addressing issues such as:** • **Why is the evaluation being undertaken?** • **How is it to be used (i.e., for learning and/or accountability functions)?** • **For whom is it undertaken?** • **Why at this particular point in time?** Score 1: The report does not describe the purpose of the evaluation. Score 2: The report describes the purpose partly, but with major gaps. Score 3: The report describes the purpose, but with minor gaps. Score 4: The report clearly describes the purpose of the evaluation.		
2.2 Evaluation object *Note: All text relevant to the quality criteria shall be assessed, irrespective of where in the report it is presented.* *Note: The report should be accessible to an outsider – even if it seems to be written as an "internal" report.*	**The report provides key information about the evaluation object.** **If the evaluation object is an intervention, or part of an intervention, this could include beneficiaries, budget, time period, geographic area, components of the intervention, expected outcomes, impact, organisational set-up/management structure/implementation arrangement and so forth.** Score 1: The report does not describe the evaluation object. Score 2: The report provides some information about the evaluation object, but with major gaps. Score 3: The report provides information about the evaluation object, but with minor gaps. Score 4: The report provides key information about the evaluation object.		
2.3 Description of the programme theory *Note: A programme theory should be described by the evaluation team even if the programme does not have an explicit programme theory.*	**The programme theory is described and appropriately assessed. This could include a discussion of the logic model, underlying assumptions and evidence base.** N/A: -9 – Not relevant. This score requires a justification from rater. Score 1: No programme theory is mentioned. Score 2: The programme theory (implicit or explicit) is only partially described and assessed, with major gaps.		

	Score 3: The programme theory (implicit or explicit) is described and assessed, but with minor gaps. Score 4: The programme theory (implicit or explicit) is well described and assessed.		
2.4 Context *Note: All text relevant to the quality criteria shall be assessed, irrespective of where in the report it is presented. Other factors than those listed here may be relevant.* *Note: The report should be accessible to an outsider.*	**The report provides relevant contextual information, including socio-economic, political and cultural factors that are significant to the object of the evaluation** Score 1: The report does not provide contextual information. Score 2: The report provides inadequate contextual information, with major gaps. Score 3: The report provides adequate contextual information, but with minor gaps. Score 4: The report provides adequate contextual information.		
2.5 Scope	**The report describes the scope of the evaluation, including temporal, geographic and other delimitations of the evaluation object.** Score 1: The report does not provide information about the scope of the evaluation Score 2: The report provides some information about the scope of the evaluation, but with major gaps. Score 3: The report describes the scope of the evaluation, but with minor gaps Score 4: The report fully describes the scope of the evaluation.		
2.6 Evaluation questions *Note: EQs (not DAC criteria) should be presented and it should be explained why they are included. EQs in TOR are "suggestions" that evaluators should comment on.*	**The report clearly describes and justifies the evaluation questions.** Score 1: No evaluation questions are mentioned. Score 2: Evaluation questions are described, but poorly justified. Score 3: Evaluation questions are described, but with minor gaps in terms of justification. Score 4: Evaluation questions are clearly described and justified.		
2.7 Existing evidence base *Note: Evaluators should consider existing evidence prior to the evaluation, the info should be presented before methods and findings chapters as it may affect eval design..*	**The report discusses the existing evidence base of relevance to the evaluation object. This may include research, previous evaluations or grey literature.** Score 1: No existing evidence-base is discussed. Score 2: The existing evidence base is poorly discussed. Score 3: The existing evidence base is discussed. Score 4: The existing evidence base is thoroughly discussed.		
Quality area 3: Methodology			
3.1 Description and justification of the evaluation design *Note: We want to know how the evaluators have chosen to approach the evaluation, and why they have chosen this specific mix of methods and data to respond to EQs.*	**The report clearly describes and justifies the overall evaluation design in terms of how the chosen package of data and methods respond to the evaluation questions.** **The description is sometimes found in an evaluation matrix.** Score 1: No explicit design is described or justified. Score 2: The overall design is partially described, but with major gaps in terms of description and/or justification. Score 3: The overall design is described and justified, but with minor gaps in terms of description and/or justification. Score 4: The overall design is clearly described and justified.		
3.2 Description of methods *Note: "Select data" refers to methods for selecting, sampling etc. sources of information, e.g. which documents to read, who to interview and where to visit.*	**The report clearly describes methods for how to select, collect and analyze data.** Score 1: Methods are not described. Score 2: Methods to sample, collect and analyze data are poorly described with major gaps Score 3: Methods to sample, collect and analyze data are adequately described, but with minor gaps. Score 4: Methods to sample, collect and analyze data are clearly described.		
3.3 Methodological application *Note: If it is not possible to figure out from the text how or if methods have been applied, score 1.*	**The application of methods for sampling/selection of sources, collection and analysis results in valid and reliable data. Poor application could involve poor or incorrect use of methods.** Score 1: Methods are not applied in a way that results in valid and reliable data. Score 2: Methods are poorly applied, with major gaps. Score 3: Methods are adequately applied, but with minor gaps. Score 4: Methods are well applied and results in valid and reliable data.		
3.4 Reliability and validity of evidence	**The report discusses the validity and reliability of evidence, i.e. discusses whether evidence can be trusted.** **Evidence here refers to data and findings derived from the collection and analysis of data.**		

	Score 1: The reliability and validity of evidence is not discussed. Score 2: The reliability and validity of evidence is poorly discussed. Score 3: The reliability and validity of evidence is partially discussed with minor gaps. Score 4: The reliability and validity of evidence is thoroughly discussed.		
3.5 Sources of evidence	**The source of evidence for all data/findings is clearly referenced throughout the report. Reference is made to documents, interviews, administrative data, literature, analysis of data, etc.** Score 1: There are no references to sources of evidence. Score 2: Sources of evidence are poorly referenced with many gaps Score 3: Sources of evidence are referenced, but with minor gaps. Score 4: Sources of evidence are clearly referenced.		
3.6 Limitations *Note: Refers to limitations **caused by** the choice of approach and methods, and how these affected the team's ability to respond to EQs.*	**The report provides a good description of the limitations arising from the chosen evaluation design.** Score 1: Limitations are not described Score 2: Limitations are poorly described. Score 3: Limitations are described but there are some gaps. Score 4: Limitations are well described.		
3.7 Ethics *Note: Refers to any ethical issues **arising from** the evaluation design. Limitations arising from ethical issues are not included here. See Norad guidelines for evals for definition of ethical issues.*	**Ethical issues arising from the evaluation and accompanying safeguards are well described.** Score 1: Ethical issues are not described. Score 2: Description of ethical issues or safeguard is poor. Score 3: Description of ethical issues and safeguards is adequate, with minor gaps. Score 4: Ethical issues arising from the evaluation and accompanying safeguards are well described.		
Quality area 4: Application of International evaluation criteria This section covers the International Evaluation criteria and describe which of the evaluation criteria and cross-cutting issues that have/has been assessed. This is relevant only to the extent that the evaluation applies one or more of the criteria or assesses cross-cutting issues. *Note: Application refers to whether an evaluation criteria or cross-cutting issue has been applied, assessed, treated, considered etc. If it is mentioned, but does not seem to have been assessed or considered, score N.*			
International Evaluation Criteria assessed in the evaluation report			
4.1 Relevance	Y/N		
4.2 Coherence	Y/N		
4.3 Effectiveness	Y/N		
4.4 Efficiency	Y/N		
4.5 Sustainability	Y/N		
4.6 Impact	Y/N		
Cross-cutting issues assessed in the evaluation report			
4.7 Human rights	Y/N		
4.8 Gender equality and women's rights	Y/N		
4.9 Climate and environment	Y/N		
4.10 Anti-corruption	Y/N		
Quality area 5: analysis, data, findings, conclusions, lessons learned and recommendations			
5.1 Response to evaluation questions *Note: If EQs are not stated and can't be found, score N/A.*	**The report fully responds to evaluation questions.** Score N/A: Not applicable as no evaluation questions could be found Score 1: The report does not respond to evaluation questions Score 2: The report partially responds to evaluation questions, but with many gaps. Score 3: The report responds to evaluation questions, but with minor gaps. Score 4: The report provides clear responses to all evaluation questions		
5.2 Findings	**Findings are founded on evidence, either directly or is derived from an analysis of evidence.** Score 1: Findings are not founded on evidence. Score 2: Findings are partially founded on evidence, but with major gaps. Score 3: Findings are founded on evidence, but with minor gaps. Score 4: Findings are clearly founded on evidence.		
5.3 Conclusions *Note: Conclusions may go beyond responding to EQs, but our main interest here is*	**Conclusions to evaluation questions flow clearly and logically from the analysis of findings.** Score 1: Conclusions to evaluation questions are not derived from the analysis of findings, or no conclusions are provided.		

if responses to evaluation questions are based on analysis of findings.	Score 2: Conclusions to evaluation questions are partially derived from the analysis of findings Score 3: Conclusions to evaluation questions are derived from the analysis of findings, but with minor gaps. Score 4: Conclusions to evaluation questions flow clearly and logically from the analysis of findings.		
5.4 Recommendations are based on conclusions *Note: The score here refers only to if recommendations are based on conclusions, even if conclusions were not founded on findings. If so, add a comment about this.*	**Recommendations are based on conclusions, options are clearly stated and discussed, and uncertainty regarding possible consequences is acknowledged.** N/A: -9: The evaluation was not mandated to provide recommendations Score 1: Recommendations are not based on conclusions. Score 2: Recommendations are only weakly based on conclusions in the report. Score 3: Recommendations are based on conclusions in the report, but with gaps in terms of discussing options and or acknowledging uncertainty regarding options. Score 4: Recommendations are based on conclusions, options are clearly stated and discussed, and uncertainty regarding possible consequences is acknowledged.		
5.5 Recommendations respond to the purpose of the evaluation	**Recommendations clearly respond to the purpose of the evaluation.** N/A: -9: The evaluation was not mandated to provide recommendations Score 1: Recommendations do not respond to the purpose of the evaluation. Score 2: Recommendations respond weakly to the purpose. Score 3: Recommendations respond largely to the purpose, but with some gaps. Score 4: Recommendations clearly respond to the purpose of the evaluation.		
5.6 Recommendations are clear and actionable *Note: If in doubt about "actionable", check OECD-DAC.* *Note: If actionable but not clear, score 2.*	**The report contains clear and actionable (targeted, timed and prioritized) recommendations.** N/A: -9, The evaluation was not mandated to provide recommendations Score 1: There are no recommendations. Score 2: Recommendations are not clear. Score 3: Recommendations are clear, but not actionable. Score 4: Recommendations are clear and actionable		
Overall quality and comments			
6.1 Overall quality of the report	**Based on the previous criteria and on an overall assessment, the report is judged to be of good quality.** Score 1: The overall quality of the report is poor. Score 2: The overall quality of the report is inadequate with major gaps. Score 3: The overall quality of the report is adequate, but with minor gaps. Score 4: The overall quality of the report is good.		
General reflections on the review	General reflections on the evaluation/review, key things missing from the report, good practise identified, positive outliers, etc.		
Below, please provide information about the findings/conclusions, lessons learned and recommendations in the report:			
Main findings/ conclusion	Main findings identified in the review, highlighting findings of particular interest and/or beyond project/programme level.		
Lessons learned	Lessons learned (of general interest) identified in the review (if any).		
Recommendations	Recommendations made in the review, that go beyond programme level (if any).		
Comments about the scoring process			
Scoring process	Reflections on the tools used in our assessment (the scoring templates), useful tips, comments, questions etc.		

Revised scoring protocol format for terms of references

General info	Report Id (as in document name, e.g. 1805RT)		
	Name of assessor (initials, e.g. IT)		
	Date of assessment (yymmdd)		
	Commissioner's reference number, if any		
	Time spent, approx. hours		
	Type of report (mid, end, etc.)		
	Report was also reviewed (Y/N)		
Key quality areas	**Quality statement and scoring guidance**	**Score**	**Comment**
1. Evaluation purpose, objectives, object and scope			
1.1 Rationale and purpose of the evaluation	**The rationale, purpose, intended users and intended use of the evaluation are stated clearly, addressing issues such as:** • **Why is the evaluation being undertaken?** • **How is it to be used (i.e., for learning and/or accountability functions)?** • **For whom is it undertaken?** • **Why at this particular point in time?** Score 1: The TOR does not describe the purpose of the evaluation. Score 2: The TOR describes the purpose partly, but with major gaps. Score 3: The TOR describes the purpose, but with minor gaps Score 4: The TOR clearly describes the purpose of the evaluation.		
1.2 Specific objectives of the evaluation	**The specific objectives of the evaluation clarify what the evaluation aims to find out.** Score 1: The specific objectives do not clarify what the evaluation aims to find out. Score 2: The specific objectives partly clarify what the evaluation aims to find out, but with major gaps. Score 3: The specific objectives clarify what the evaluation aims to find out, but with minor gaps. Score 4: The specific objectives fully clarify what the evaluation aims to find out.		
1.3 Context of the development intervention being evaluated	**The TOR provides relevant contextual information, including socio-economic, political and cultural factors that are significant to the object of the evaluation** Score 1: The TOR does not provide contextual information. Score 2: The TOR provides inadequate contextual information, with major gaps. Score 3: The TOR provides adequate contextual information, but with minor gaps. Score 4: The TOR provides adequate contextual information.		
1.4 Previous evaluations	**The ToR states whether previous evaluations exist, and If applicable, identifies relevant issues.** Score 1: No previous evaluations are mentioned. Score 2: Previous evaluations are poorly discussed. Score 3: Previous evaluations are discussed. Score 4: Previous evaluations are thoroughly discussed.		
1.5 Evaluation object	**The TOR provides key information about the evaluation object.** **If the evaluation object is an intervention, or part of an intervention, this could include beneficiaries, budget, time period, geographic area, components of the intervention, expected outcomes, impact, organisational set-up/management structure/implementation arrangement and so forth.** Score 1: The TOR does not describe the evaluation object. Score 2: The TOR provides some information about the evaluation object, but with major gaps. Score 3: The TOR provides information about the evaluation object, but with minor gaps. Score 4: The TOR provides key information about the evaluation object.		
1.6 Scope	The TOR describes the scope of the evaluation, including temporal, geographic and other delimitations of the evaluation object. Score 1: The TOR does not provide information about the scope of the evaluation Score 2: The TOR provides some information about the scope of the evaluation, but with major gaps. Score 3: The TOR describes the scope of the evaluation, but with minor gaps Score 4: The TOR fully describes the scope of the evaluation.		
1.7 Evaluation criteria	**Based on the evaluation mandate, the TOR identifies the relevant criteria (DAC criteria and cross-cutting issues) for the evaluation:** Score 1: Evaluation criteria and issues are not identified. Score 2: Evaluation criteria and issues are identified, but with major shortcomings regarding relevance.		

	Score 3: Evaluation criteria and issues are identified, but with minor shortcomings regarding relevance. Score 4: Evaluation criteria and issues are clearly identified and relevant to the evaluation mandate.		
1.7a Criteria and issues covered	**International Evaluation Criteria covered** This section refers to if the International Evaluation criteria and cross-cutting issues have/has been included in the TOR.		
1.7.1 Relevance	Y/N		
1.7.2 Coherence	Y/N		
1.7.3 Effectiveness	Y/N		
1.7.4 Efficiency	Y/N		
1.7.5 Sustainability	Y/N		
1.7.6 Impact	Y/N		
1.8 Evaluation questions	**The evaluation questions are customized and rendered specific to users' (as defined in the rationale and purpose section) information needs.** Score 1: No evaluation questions are mentioned. Score 2: Evaluation questions are described, but with major gaps in terms of relevance and customization. Score 3: Evaluation questions are described, but with minor gaps in terms of relevance and customization. Score 4: Evaluation questions are clearly described, relevant and customized.		
1.9 Feasibility	**The scope of work proposed by the TOR is feasible given the timeframe and resources provided. The TOR contain a limited/prioritized number of evaluation questions that are clear and relevant to the object and purpose of the evaluation.** Score N/A: The TOR does not provide sufficient information to assess feasibility. Score 1: The scope of work is not feasible given the timeframe and resources provided. Score 2: The scope of work is feasible given the timeframe and resources provided, but with major difficulties. Score 3: The scope of work is feasible given the timeframe and resources provided, but with minor difficulties. Score 4: The scope of work is fully feasible given the timeframe and resources provided.		
2. Review Process and QA			
2.1 Review process	**The ToR clearly explains what is expected of the Consultant in terms of:** • **Required evaluation phases (e.g. having an inception phase).** • **Demands regarding data collection and validation.** • **Instructions for preparing the evaluation report.** • **Roles and responsibilities of the evaluation team, the commissioner and other involved parties.** Score 1: Evaluation process and roles are not described. Score 2: Evaluation process and roles are poorly described with many shortcomings in explanations and/or appropriateness. Score 3: Evaluation process and roles are described, but with minor shortcomings in explanations and/or appropriateness. Score 4: Evaluation process and roles are clearly explained and are appropriate.		
2.2 Deliverables	**The TOR identifies the mandatory deliverables and milestones. This could include:** **· inception report (if applicable)** **· debriefing / validation sessions** **· draft and final evaluation report** **· presentation of the report (optional)** **The schedule identifies the key phases of the evaluation.** Score 1: Deliverables and milestones are not identified. Score 2: Deliverables and milestones are identified, but with major gaps. Score 3: Deliverables and milestones are identified, but with minor gaps. Score 4: Deliverables and milestones are clearly identified.		
2.3 Quality assurance	**The TOR specifies the required quality assurance mechanisms including that the evaluation will follow professional norms and standards and OECD-DAC guidelines.** Score 1: Quality assurance requirements are not specified. Score 2: Quality assurance requirements are specified, but with major gaps. Score 3: Quality assurance requirements are specified, but with minor gaps. Score 4: Quality assurance requirements are clearly specified.		
3. Overarching and cross-cutting criteria			

3.X Human rights	**Human rights are reflected in the TOR where appropriate (context, design, questions around effectiveness and impact).** Score 1: Human rights are not reflected, even if appropriate. Score 2: Human rights are reflected, but with major gaps. Score 3: Human rights are reflected, but with minor gaps. Score 4: Human rights are well reflected.		
3.1 Gender	**Gender dimensions and women's rights are explicitly addressed in all relevant parts of the TORs (context, questions, approach, design, methods, team composition).** Score 1: Gender dimensions and women's rights are not reflected. Score 2: Gender dimensions and women's rights are reflected, but with major gaps. Score 3: Gender dimensions and women's rights are reflected, but with minor gaps. Score 4: Gender dimensions and women's rights are well reflected.		
3.2 Climate and Environment	**Climate and environment dimensions are reflected in the TOR where appropriate (context, design, questions around effectiveness and impact).** Score 1: Climate and environment dimensions are not reflected, even if appropriate. Score 2: Climate and environment dimensions are reflected, but with major gaps. Score 3: Climate and environment dimensions are reflected, but with minor gaps. Score 4: Climate and environment dimensions are well reflected.		
3.3 Anti-corruption	**Anti-corruption issues are reflected in the TOR (e.g as part of risks or context).** Score 1: Anti-corruption issues are not reflected. Score 2: Anti-corruption issues are reflected, but with major gaps. Score 3: Anti-corruption issues are reflected, but with minor gaps. Score 4: Anti-corruption issues are well reflected.		
3.4 Ethics	**Ethical considerations (consent, protection, participation, independence) and requirements are explicitly addressed in the TOR.** Score 1: Ethical considerations and requirements are not reflected. Score 2: Ethical considerations and requirements are reflected, but with major gaps. Score 3: Ethical considerations and requirements are reflected, but with minor gaps. Score 4: Ethical considerations and requirements are well reflected.		
3.5 Expected limitations to the review	**Expected limitations to the evaluation are identified (methods, sources of info, disaggregated data, time, budget).** Score 1: Expected limitations are not identified. Score 2: Expected limitations are identified, but with major gaps. Score 3: Expected limitations are identified, but with minor gaps. Score 4: Expected limitations are clearly identified.		
4 OVERALL RATING			
4.1 Overall rating of the ToRs	**The TOR provides a sound basis for the evaluation, that will guide the evaluation manager and evaluation team on how to effectively fulfill the objectives of the evaluation.** Score 1: The TOR does not provide a sound basis for the evaluation. Score 2: The TOR provides a sound basis for the evaluation, but with major gaps. Score 3: The TOR provides a sound basis for the evaluation, but with minor gaps. Score 4: The TOR provides a sound basis for the evaluation.		
Good practise?			
General comments			

Annex 4: Decentralised evaluations included in the assessment

Note: Question marks indicate missing or uncertain information.

Table 4: Quality assessed decentralised evaluations

Title of the report	Year	Accessed from	Implemented by	Commissioner
Review of the Joint Programme for the Kigoma Region (JPK)	2020	Royal Norwegian Embassy in Dar es Salaam	Norad + external	Embassy
End review of Energy farm in Ukraine	2020	Department for Climate, Energy and Environment	LTS International Limited	Norad
Provision of adequate tree seed portfolios (PATSPO)	2020	Royal Norwegian Embassy in Addis Ababa	Independent consultants	Embassy
Desk Review: Project Performance of Food Security and Adaptation to Climate Change in Rural Niger	2020	Royal Norwegian Embassy in Bamako	LTS International Limited (UK) + Norad	Embassy
External Review of the ongoing projects in Montenegro, North Macedonia and Serbia supported by Norway and Sweden and implemented by UNOPS	2020	Royal Norwegian Embassy in Belgrade	Independent consultant	Embassy
End Review of Technical Cooperation for Development of Kikuletwa Power Station as a Hydropower Training Centre and for Electricity Generation	2020	Royal Norwegian Embassy in Dar es Salaam	Norconsult	Embassy
Near-End Review of the Oil for Development Programme in Tanzania Phase II	2020	Royal Norwegian Embassy in Dar es Salaam	Norad	Embassy
End Review of the Agricultural Council of Tanzania's Agricultural Partnership Programme Phase II (TAP II) and an Appraisal of the Partnership for Scale Programme (PFS)	2020	Royal Norwegian Embassy in Dar es Salaam	LTS International Limited	Norad
Review - FAO's Emergency Livelihoods Response Program	2020	Royal Norwegian Embassy in Juba	Norad	Embassy
The Malawi Parliament Enhancement Project Mid-Term Review Report	2020	Royal Norwegian Embassy in Lilongwe	Independent	Embassy
Midterm Review (MTR) of Agreement RAF 17/0053 – High-Level Mediation and Negotiations Training in Africa Between The Royal Norwegian in Pretoria and The Centre for Mediation in Africa at The University of Pretoria	2020	Royal Norwegian Embassy in Pretoria	Independent	Embassy + partner
End-term evaluation of Project RSA-3005, RAF-16/0046 Women Pioneers in The Judiciary Supporting Female Judges and Female Law Students	2020	Royal Norwegian Embassy in Pretoria	Independent	Embassy + partner
Training the front-line officers for better combat of fisheries crime. A mid-term review of the FishFORCE project at Nelson Mandela University, South Africa.	2020	Royal Norwegian Embassy in Pretoria	Norwegian College of Fishery Science, University of Tromsø	Embassy
Report of the review of the institute for security studies (ISS)	2020	Royal Norwegian Embassy in Pretoria	Independent	Embassy + partner
Mid-Term Review: Peacebuilding in Africa project, In Transformation Initiative (ITI)	2020	Royal Norwegian Embassy in Pretoria	Independent	Embassy + partner?
Review of the project Reducing the impact of Large-Scale Agricultural Investments in the Mekong Region on Communities, Forest and Climate Change"	2020	Department for Climate, Energy and Environment	LTS International Limited	Embassy
Guyana MRV Support – Mid Term Evaluation	2020	Evalueringsportalen and Department for Climate, Energy and Environment	LTS International Limited	Embassy
NORAD and NICFI supported project: Advancing Jurisdictional Programs for REDD+ and Low-	2020	Department for Climate, Energy and Environment	KPMG	Norad

Emissions Development: Governors' Climate & Forests Task Force (GCFTF) DRAFT				
Education and Schools in Afghanistan: Mid-term review of AFG-14/0022 Equitable Access to Quality Education in Faryab, Afghanistan	2020	Section for South Asia and Afghanistan	NCG + Tadbeer Consulting	MFA
End review Rikskonsertene/ Kulturtanken's contract with the royal norwegian embassyin india 2008-2018	2020	Royal Norwegian Embassy in New Delhi	Independent	Embassy
End-Review of the project Management of Catastrophic Disasters in Uttarakhand	2019	Norad Section for Food Security and Environment	Norad + Ecorys	Embassy
Gjennomgang av Visjon 2030-mekanismen	2019	The Knowledge Bank	KPMG	Norad
Strategic Support to Food Security and the Agricultural Sector in Malawi: Final Report, NORAD Call-Down 12	2019	Royal Norwegian Embassy in Lilongwe	LTS International Ltd	Norad
Fortaleciendo las Capacidades Jurisdiccionales de la Corte Interamericana de Derechos Humanos y de la Difusión de su Trabajo	2019	Royal Norwegian Embassy in Mexico	Just Governance Group Ltd.	Embassy
Adapting agriculture to climate change: collecting, protecting and preparing crop wild relatives	2019	www.norad.no		Norad + partner?

Annex 5: Data and descriptive statistics

Table 5: Data and descriptive statistics, terms of reference quality criteria

Terms of reference quality criteria	Average	Standard deviation	Score 1	Score 2	Score 3	Score 4	Score n/a	No of reports
1.1 Rationale and purpose of the evaluation	3.50	0.76	0	4	4	16	0	24
1.2 Specific objectives of the evaluation	3.75	0.66	1	0	3	20	0	24
1.3 Context of the development intervention being evaluated	2.08	1.04	9	7	5	3	0	24
1.4 Previous evaluation	1.79	1.12	14	5	1	4	0	24
1.5 Evaluation object	3.25	0.78	1	2	11	10	0	24
1.6 Scope	3.29	0.79	0	5	7	12	0	24
1.7 Evaluation criteria	3.00	0.87	1	6	9	8	0	24
1.8 Evaluation questions	3.42	0.70	1	0	11	12	0	24
1.9 Feasibility	3.13	0.70	0	3	8	5	8	24
2.1 Review process	2.92	0.86	1	7	9	7	0	24
2.2 Deliverables	3.25	0.88	1	4	7	12	0	24
2.3 Quality assurance	1.42	0.81	18	3	2	1	0	24
3.X Human rights	2.00	1.08	12	2	8	2	0	24
3.1 Gender	2.04	1.02	10	5	7	2	0	24
3.2 Climate and Environment	2.00	1.12	12	3	6	3	0	24
3.3 Anti-corruption	1.63	0.95	15	5	2	2	0	24
3.4 Ethics	1.46	1.00	19	2	0	3	0	24
3.5 Expected limitations to the review	1.42	0.64	16	6	2	0	0	24
4.1 Overall rating of the ToRs	2.67	0.55	0	9	14	1	0	24

Table 6: Evaluation criteria covered in terms of references

Evaluation criteria	Percent included	Included	Not included
1.7.1 Relevance	92%	22	2
1.7.2 Coherence	46%	11	13
1.7.3 Effectiveness	96%	23	1
1.7.4 Efficiency	88%	21	3
1.7.5 Sustainability	71%	17	7
1.7.6 Impact	71%	17	7

Table 7: Data and descriptive statistics, evaluation report quality criteria

Report quality criteria	Average	Standard deviation	Score 1	Score 2	Score 3	Score 4	Score N/A	No of reports
1.1 Executive summary	2.48	0.79	2	13	9	3	0	27
1.2 Style and structure	2.89	0.79	0	10	10	7	0	27
2.1 Purpose of the evaluation	2.96	0.96	3	4	11	9	0	27
2.2 Evaluation object	2.93	0.77	0	9	11	7	0	27
2.3 Description of the programme theory	2.52	1.03	4	12	4	7	0	27
2.4 Context	2.70	0.85	1	12	8	6	0	27
2.5 Scope	2.85	0.93	3	5	12	7	0	27
2.6 Evaluation questions	2.07	0.94	9	9	7	2	0	27
2.7 Existing evidence base	2.04	0.84	7	14	4	2	0	27
3.1 Description and justification of the evaluation design	2.44	1.03	5	11	5	6	0	27
3.2 Description of methods	2.56	0.83	2	12	9	4	0	27
3.3 Methodological application	2.41	1.03	7	6	10	4	0	27
3.4 Reliability and validity of evidence	1.93	1.05	13	6	5	3	0	27
3.5 Sources of evidence	2.48	0.63	0	16	9	2	0	27
3.6 Limitations	1.96	1.17	14	5	3	5	0	27
3.7 Ethical issues	1.37	0.91	23	0	2	2	0	27
5.1 Response to evaluation questions	3.04	0.82	0	8	8	9	2	27
5.2 Findings	2.67	0.77	0	14	8	5	0	27
5.3 Conclusions	3.00	0.86	0	10	7	10	0	27
5.4 Recommendations are based on conclusions	2.96	0.82	1	6	11	7	2	27
5.5 Recommendations respond to the purpose of the evaluation	3.28	0.87	0	7	4	14	2	27
5.6 Recommendations are clear and actionable	2.72	0.72	0	11	10	4	2	27
6.1 Overall quality of the report	2.74	0.75	1	9	13	4	0	27

Table 8: Evaluation criteria included, evaluation reports

Evaluation criteria	Percent included	Included	Not Included	No of reports
4.1 Relevance	96%	26	1	27
4.2 Coherence	67%	18	9	27
4.3 Effectiveness	100%	27	0	27
4.4 Efficiency	93%	25	2	27
4.5 Sustainability	85%	23	4	27
4.6 Impact	93%	25	2	27
4.7 Human rights	56%	15	12	27
4.8 Gender equality and women's rights	74%	20	7	27
4.9 Climate and environment	63%	17	10	27
4.10 Anti-corruption	44%	12	15	27

Figure 1: Distribution of scores for terms of references

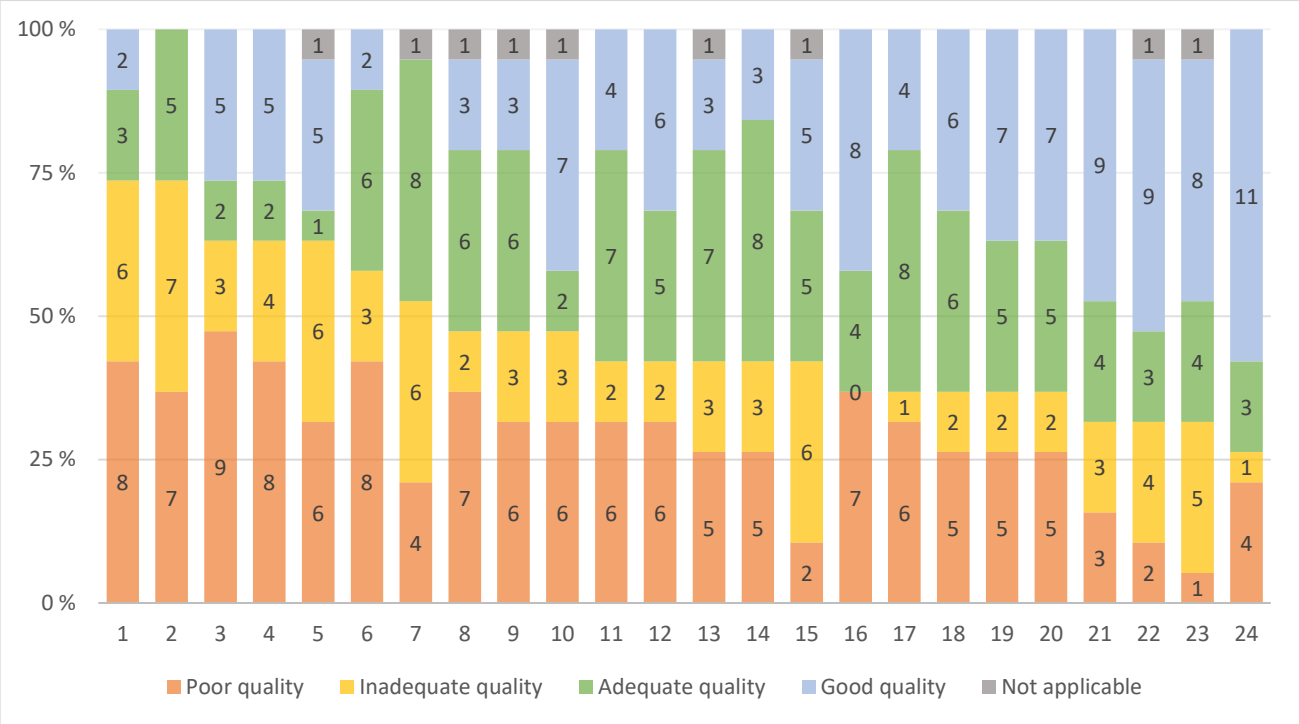


Figure 2: Distibution of scores for terms of reference quality criteria

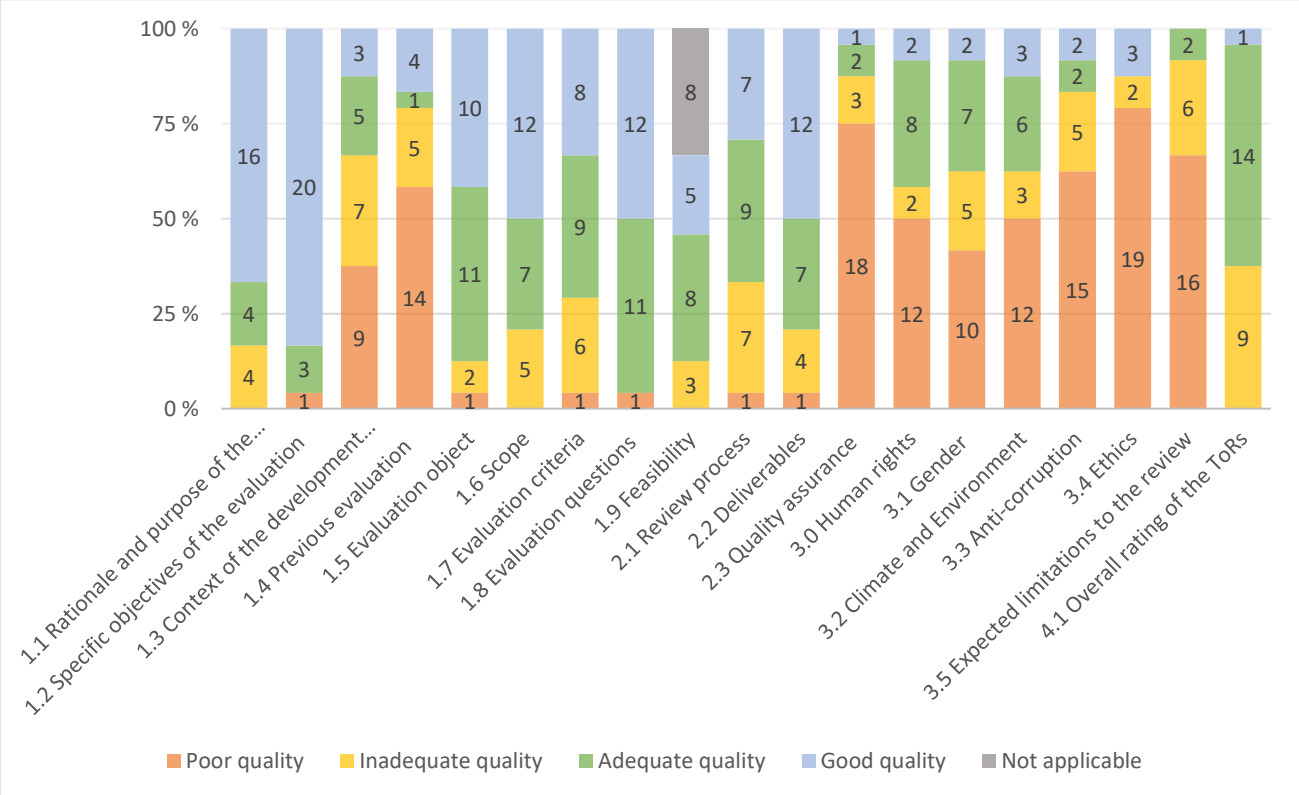


Figure 3: Average scores for terms of reference quality criteria

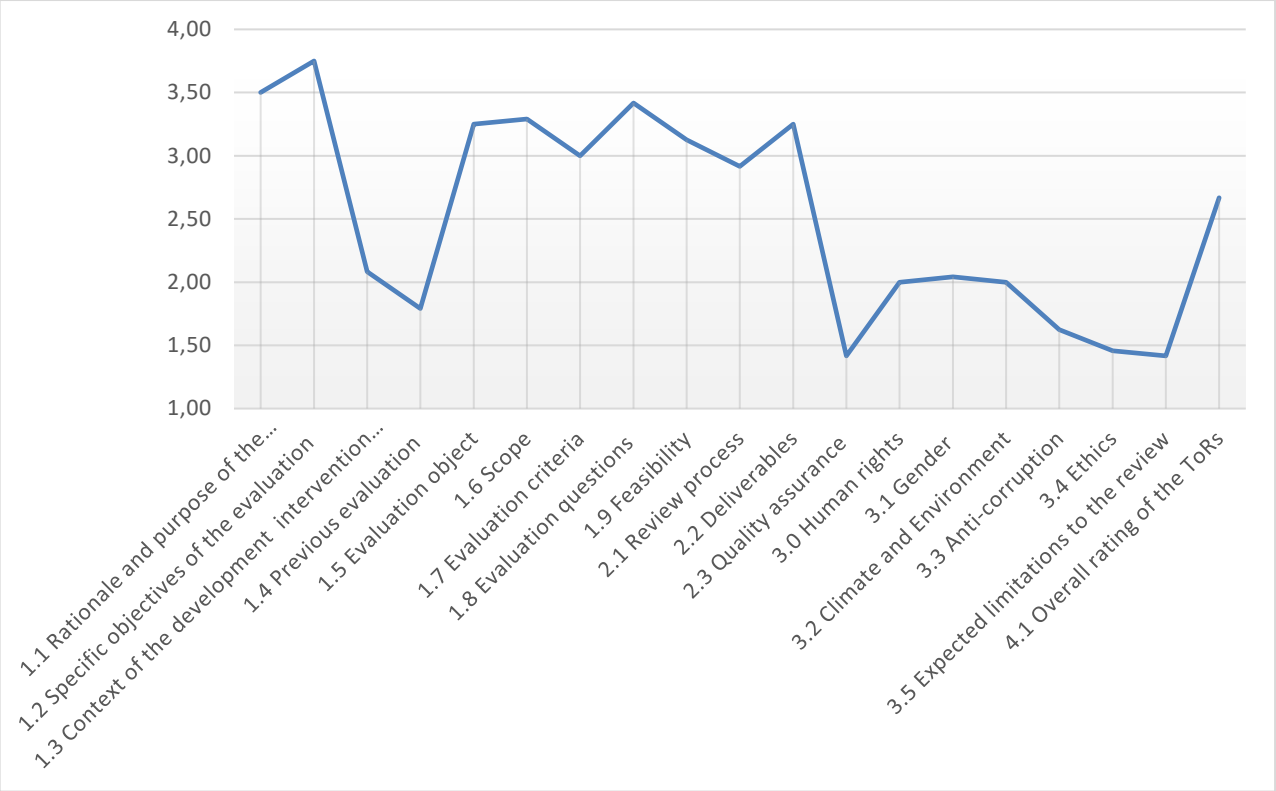


Figure 4: Distibution of scores for evaluation reports

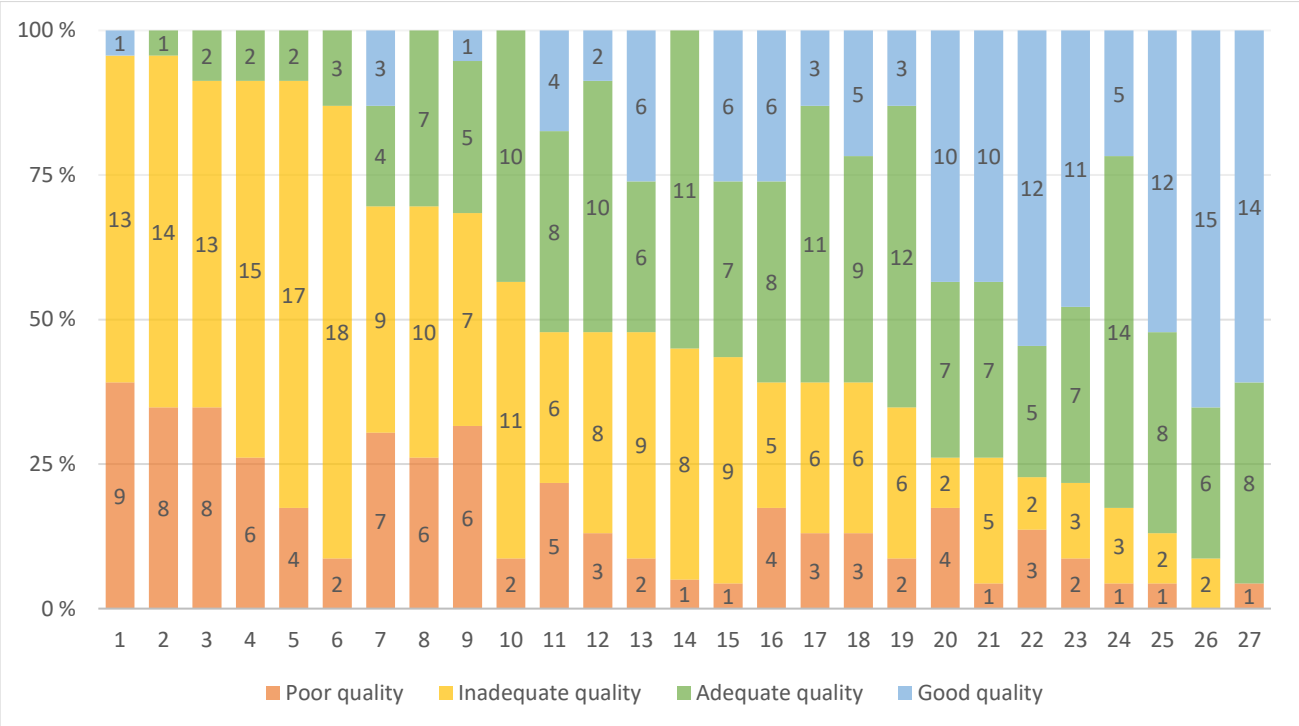


Figure 5: Distribution of scores for report quality criteria

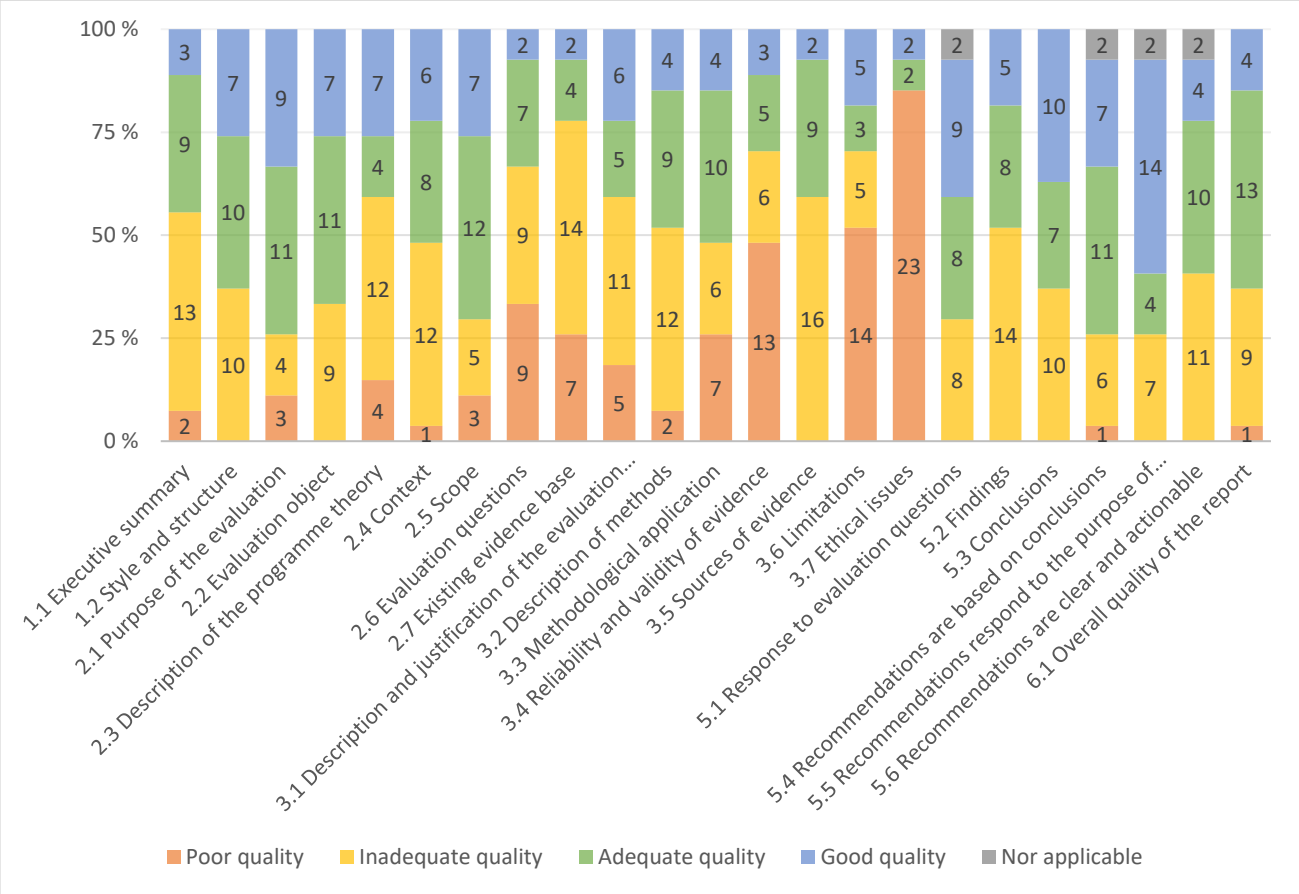
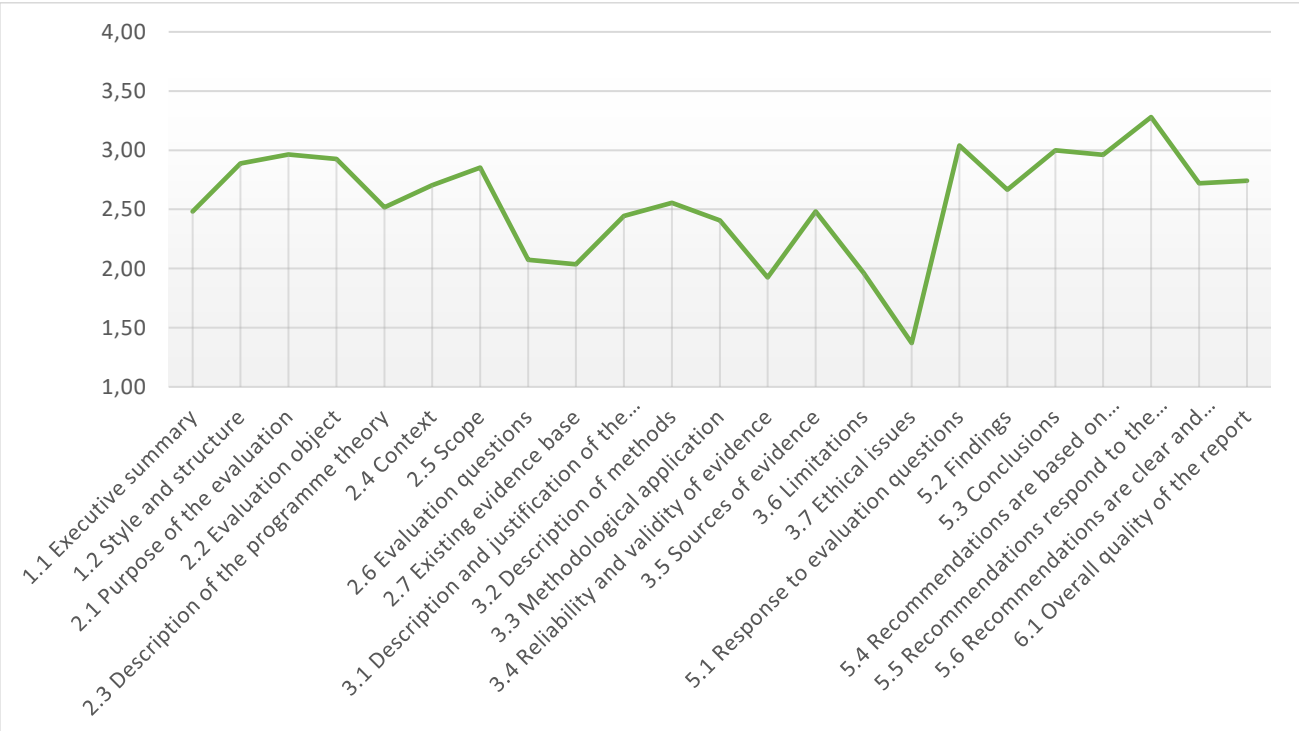


Figure 6: Average scores for report quality criteria



Annex 6: Good practise evaluations

Annex 6.1 End review of the Tanzanian Agricultural Partnership programme phase II

End Review of the Agricultural Council of Tanzania's Agricultural Partnership Programme Phase II (TAP II) and an Appraisal of the Partnership for Scale Programme (PFS)[1]

The Agricultural Council of Tanzania (ACT) is a private sector apex organisation that aims to bring together all the country's stakeholders involved in agriculture. ACT and Norway have been partners since 2006, and Norway has supported the Tanzanian Agricultural Partnership (TAP) since 2008. TAP Phase II (2014-20) had the overall objective of stimulating economic growth through commercially oriented activities and investments, explicitly through increased investments in the agricultural sector. Its specific objectives were to:

- increase smallholder farmers' agricultural productivity and profitability to enhance food security and incomes
- enable improvement in the agri-business environment within a value chain framework.

The evaluation was a combination of an end review of TAP Phase II and style and an appraisal of a proposal for TAP Phase III. The report is clearly structured and well written, although the headings are a bit confusing: Findings of the evaluation are presented in chapters referred to as "results" and arranged along the programme's intended outcomes, and a chapter entitled "Major Findings" is in fact a conclusions chapter. Once you realise this, these do not create any confusion or problems as far as understanding the report. Another oddity, though, is that the author and evaluation team are not identified.

Purpose, rationale, use and users are clearly stated or easily understood from the text. The scope of the evaluation is fairly clear, but more could have been said about geographical areas and programme components covered. The object of the evaluation is well described. The theory of change is briefly summarised in a table, and the intended outcomes are systematically used as section titles for accounting of results. Activities, outputs and indicators are less consistently described. A good description of context is provided, both regarding agriculture in general in Tanzania and regarding the project environment, but more could have been said about the target groups.

Evaluation questions are not listed in the report but are available in the annexed terms of reference. They are indirectly presented, however, by the way their answers are presented within the sections for intended outcomes, cross-cutting issues and other aspects to be assessed, like handling of risks.

An earlier mid-term review is mentioned and extensively used in a compilation of recommendations from that report that were subsequently adopted during implementation. Such systematic follow-up of lessons raised in the mid-term review is unusual in the reports the team reviewed. However, almost nothing is said about other existing studies in the field that may have been relevant for the project design and/or the focus of the review.

The overall design of the evaluation is briefly but well described and justified. The report describes methods appropriately, including brief descriptions of selection, collection and analysis of information, as can be seen

[1] LTS International (2020). End Review of the Agricultural Council of Tanzania's Agricultural Partnership Programme Phase II (TAP II) and an Appraisal of the Partnership for Scale Programme (PFS). Norad.

in the passages quoted below from the evaluation. Interview questions, including a presentation of the review and interview process to key informants, are annexed to the report. Sources and triangulation are frequently referred to, indicating that methods have been applied as intended. The following text from the report illustrates how overall design and methods were described:

"To achieve the mission objectives within the time available a rapid participatory and consultative approach was used with ACT staff, partners and key observers (i.e. local and regional authorities, relevant government departments and development partners) to ensure that their experience and knowledge fully inform the evaluation and appraisal. Given the scope of the work (see the ToR in Annex 1) in terms of the number of questions and time available to address each of these, coverage was broad as opposed to deep. Answers are perceived a sufficiently well-grounded in available evidence to be useful in the adjustment of the PFS proposal and for the Grant Manager's follow-up of the project.

The approach was based on (i) document review (project document with annexes, reviews (mid-term and end-term,) of previous project phases, appraisals and projects documents and reports) and (ii) key informant interviews (KIIs) and group discussions (GDs). A list of documents reviewed by the mission is included in Annex 2: Document List. A list of people contacted, and interviews is included in Annex 3: List of Contacts. All interviews were performed using online communication as the Covid-19 pandemic reduced the ability to travel and perform face to face meetings. The review team did not engage physically or through online or phone-based communication at a farm level. KIIs and DGs were based on a list of potential interviewees suggested by ACT, Royal Norwegian Embassy (RNE) and NORAD. Interviews were held with people that were available during the data collection period. […] Analysis compared and contrasted findings to develop common themes following the mission's key questions. These findings and themes were developed into a draft report. Following data collection an online debriefing was held with NORAD and ACT staff. Report generation was iterative with drafts produced for comment that were responded to and a final report developed".[2]

The report contains a very good discussion of limitations to the data that were collected: There is a clear account of what data are missing or doubtful in the project reporting and what categories of key informants were not possible to reach (mainly from private sector). This is illustrated in the following text from the report:However, the consequences for how to interpret evaluation findings could be even more clearly expressed.

"A limitation of the study was the underrepresentation of private sector representatives in interviews. These important actors were more difficult to contact and less willing to respond to invitations for interview. Time and budget available for the review constrained the level of follow-up that was possible.

The review and appraisal make significant use of ACT reporting […] These reports do not describe the methods used to collect data and generate their findings. Therefore, it is hard to assess their quality, particularly in relation to outcome level reporting […].

Selection bias was experienced by the study. It is suggested that the most successful, active, responsive, or engaged beneficiaries were contacted, in particular at a district level. The mission mitigated this risk by using triangulation and cross checking as much as possible. Ideally the study would have also engaged districts that had been less active under TAP II.

The rapid approach used by the review limited the depth of information that the team could gather and the amount of triangulation that could be performed. The review was limited in the level of detailed information it could collect about specific partner interventions."[3]

However, although the description of the quality of data is clear, the report does not clearly state how this affects the credibility of the evaluation findings: If there e.g. is a risk that the selection bias makes the programme seems more successful than it actually is, this should be clearly stated.

[2] LTS International, 2020, End Review of the Agricultural Council of Tanzania's Agricultural Partnership Programme Phase II (TAP II) and an Appraisal of the Partnership for Scale Programme (PFS). Norad. P. 15.
[3] Ibid.

Ethical issues are not mentioned, although the report includes a note advising that its list of contacts should not be included if the report is publicly circulated. The version of the report made available to the team of raters included this list of contacts. The report does not discuss human rights issues and while the terms of reference request an assessment of anti-corruption issues, the report states that it will not do so. Gender is discussed several times in the report and results assessed. Climate and environment issues are built into the project, in particular conservation agriculture methods, and are assessed.

All evaluation questions seem to be answered, although not as listed in the terms of reference; instead, they are responded to either under each evaluation criterion and cross-cutting issue or as results in relation to intended outcomes.

Evidence is extensively used and directly or indirectly referred to when presenting findings and/or results. A large number of footnotes not only provide sources of information but also add detail about data and comment on evidence and findings. Some parts of the report are less well referenced than others. This paragraph from the report illustrates that information from different sources has been compared and differences are clearly reported:

"Mixed results are found for the overall adoption of Good Agricultural Practice (GAP). Some observers report increases in GAP as generally occurring in Tanzania, whereas others suggest adoption has not been so encouraging. The later view is consistent with available secondary data (See Section 4.1.1)."[4]

More specific material drawn from interviews could have been used. For instance, rather than stating, as it did, that *"[S]ome observers report increases",* the report could have presented the number or share of observers that reported increases.

Conclusions seem well founded on the findings (although, as noted, the relevant chapter is entitled "Major Findings" instead of "Conclusions"). The recommendations are properly based on the conclusions and respond to the review purpose. Each recommendation is discussed, not just briefly stated; includes possible consequences; and is justified. The recommendations are clear and targeted but not timed or prioritised.

Overall, the report is clearly structured, systematic and well written. It is interesting and probably a very useful approach to combine the end review of one phase with the appraisal of the proposal for a possible new phase.

The review found that TAP Phase II has contributed to impressive increases in agricultural productivity and profitability but that these increases are localised and only being achieved by a small proportion of targeted farmers.[5] It also concludes that there is conflicting information about the effectiveness of some components of the programme, with some sources claiming better results than claimed by others. An important observation in the report was that the implementing partner apparently had a conflict of interest, as it was both pursuing its own interest to cover the whole country to organise more farmers while it was also being contracted to implement the project, resulting in doubtful implementation effectiveness. A key learning from the review is to not launch an implementation model at full scale without testing it first in a pilot project.

[4] Ibid, p. 18.
[5] Ibid, p. 7.

Annex 6.2 Mid-term review of the FishFORCE project

Training the front-line officers for better combat of fisheries crime: A mid-term review of the FishFORCE project at Nelson Mandela University, South Africa[6]

FishFORCE is a project at Nelson Mandela University that addresses fisheries crime. The core activity is training, supported by research. Training started with the front-line officers engaged in control activities primarily in the fisheries sector and has moved on to customs, police and the judiciary. The ambition is to gradually expand these activities to neighbouring countries.

The one-page summary presents a brief history of the programme, some key conclusions and recommendations. It does not provide information about the purpose, scope or evaluation questions or how the review was implemented. The report is well written, with care given to guiding the reader through the report. There is a clear structure that ensures evaluation questions are responded to and chapters are in a logical order. The report is of high quality in terms of both language and flow of information.

There is brief but clear information about the purpose, use and users of the report. The programme is described well, providing the reader a good understanding of the set-up, intended outcomes, activities and components. The programme theory (documented results framework) is described and assessed, both in the introduction and as a separate evaluation question. The programme theory is also referred to in several parts of the report when various aspects of the programme are discussed and analysed. The assessment of the programme theory could have gone a bit further in terms of assessing the possible contribution of the project toward the intended purpose, instead of dismissing the intended impact indicators as being too difficult to assess due to many other influencing factors.

The context chapter is rather long, but the information presented is relevant and useful. It also provides relevant contextual information, giving the reader a good understanding of the fisheries sector and crimes in South Africa and the region, key stakeholders, main problems, etc. Existing evidence is extensively used in the context chapter as well as in response to some of the evaluation questions.

Evaluation questions are not separately presented or justified in the report, but there is reference to the inception report that presents and elaborates them. As the findings chapter is organised by evaluation areas and questions, the evaluation questions are clear.

The evaluation process is described, but there is not a thorough discussion of why the specific methods were selected. The report refers to the terms of reference as the basis of the review and to an inception report in which the evaluation questions were elaborated. The design is described in a more narrative way than it is in many other reports:

"We started the work by reviewing the project documents listed in Appendix 2. Based on that reading, we wrote an inception report to the Embassy. The Embassy already had structured the questions to be addressed in the review. In the inception report, we elaborated more detailed questions, identified organisations and individuals that should be approached for interviews, and set up a schedule for field work in Port Elizabeth, Cape Town and Pretoria. Both the supplementary questions and the lists of potential interviewees were long, and we had to prioritise what we could reach since a midterm review is supposed to be less thorough than a final evaluation."[7]

The report notes that selection of interviewees is critical for the information obtained and describes how interviewees were identified. It also includes a description of how interviews were conducted and how the

[6] Sander, Santos and Pretorius (2020). Training the front-line officers for better combat of fisheries crime. A mid-term review of the FishFORCE project at Nelson Mandela University, South Africa. Norwegian College of Fishery Science, University of Tromsø - Norway's Arctic university.
[7] Ibid. p. 2.

analysis was done. Analysis is described, with reference to triangulation between sources and methods and discussions in the project team. Data collection tools (for example interview guides) are not presented.

Much of the analysis was done in discussions among the evaluators and there are no interview transcripts. This decreases the transparency of the analysis process. However, the way findings are presented in the report indicates that the methods selected were applied as intended, including a thorough analysis of data and comparison of information from different sources and methods.

However, if the selection of interviewees was affected to a large extent by the information received from FishFORCE, this may have contributed to biased data. Similarly, the interruption caused by the COVID-19 pandemic may also have affected data, as some stakeholders were excluded as a result. The report concludes that despite this, the data sources were sufficiently diverse and that the main limitation was caused by lack of time:

"*Due to measures taken by the Norwegian government to combat the corona virus, we unfortunately had to cancel the rest of the fieldwork and leave from Cape Town after only one visit to a fishing community and interviews in two organisations. We have tried to compensate for this with a few video and telephone interviews afterwards. The result is the 38 interviewees listed in Appendix 3, covering a wide base of expertise and links to FishFORCE. Our judgment is that this is comprehensive and cover the most relevant organisations. The major weakness is that we did not have enough time to explore issues related to the outcomes of the project in relation to its context*."[8]

Limitations arising from the design are described, clearly indicating what the evaluators have not been able to assess:

"We have based our review on the written documents referred to in Appendix 2a. It has not been possible to systematically read reports made by the project, which may have discussed certain issues more in detail. Similarly, we have not qualitatively evaluated the content of the courses and the research. Our impression on these matters are based on overviews provided, such as in Appendix 5."

The consequences of these limitations for the quality of data are, however, not described. Such consequences could, for example be that specific findings are less certain because of these limitations .

Referencing to sources is made for external evidence, interviews and project documents but is not always very clear. For example, Table 2 in the evaluation presents a results monitoring framework. But it is not clear if this is a copy of a project document or information put together by the evaluators.

Ethical issues and quality assurance are not discussed, although these are specifically requested in the terms of reference, which notes, *"Other issues to address are routines for quality assurance; ethical standards (e.g., confidentiality of informants, sensitivity and respect to stakeholders, Do No Harm, Code of conduct)."*[9] This may have been discussed in the inception report, though.

The discussion of cross-cutting issues is more extensive than in most other reports but still does not fully respond to the request in the terms of reference for assessment of negative effects and mitigation efforts. Apart from this, the report provides clear responses to evaluation questions. Findings are clearly founded on evidence by referencing to sources or to the evaluators' analysis. Conclusions are organised around the evaluation criteria. They follow clearly from findings and provide summaries of findings for each of the six evaluation criteria. The conclusion regarding impact is interesting. It shows that the report has analysed the programme theory and openly expresses a concern regarding the role of the donor in formulating results frameworks, as the following passage illustrates:

[8] Ibid. p. 3.
[9] Ibid. p. 33.

"Impact: The current results framework has identified impacts at a too high level and mostly seems like an attempt to live up to formal requirements from NORAD. FishFORCE has only a marginal influence on the selected criteria for impacts compared to other government policies and initiatives."[10]

Recommendations are based on the responses to evaluation questions provided in the findings chapter. Also presented is a set of recommendations that go beyond the programme; this is clearly stated, and it is also stated that these recommendations are not built on findings presented in the report.

The recommendations respond to the purpose of the evaluation and the authors have clearly aimed to deliver what the terms of reference asked for. However, there could have been more discussion of options and uncertainty, and the recommendations could be more clearly formulated. Some are expressed as suggestions and things that should or could be achieved rather than in terms of how the recommendations should be achieved: *"Experience with co-management, information campaigns and communal involvement in compliance monitoring should be gained."*[11]

[10] Ibid. p. 25.
[11] Ibid. p. 27.

Annex 6.3: End-term evaluation of the Women Pioneers in the Judiciary project

End-term evaluation of Project RSA-3005, RAF-16/0046: Women Pioneers in The Judiciary - Supporting Female Judges and Female Law Students[12]

This project was implemented by the Democratic Governance and Rights Unit (DGRU) at the University of Cape Town in partnership with the South African Chapter of the International Association of Women Judges (SAC-IAWJ). The Norwegian embassy has supported the project since 2017 with a total of NOK 3 200 000 over a three-year period (2017-20).

The overall goal of the project is to have a more equal gender distribution in the judiciary and legal profession, with more women available to apply for jobs in the judiciary. Linked to this, the aim was to expose ordinary citizens to a more representative and diverse bench and profession.[13]

The project has four intended outcomes:

- Outcome 1: Women get practical experience through mentorship and internships
- Outcome 2: Women judges receive quality research to help them write better judgements
- Outcome 3: The SA-IAWJ begins to develop a track record of doing relevant gender work in the broader community and thus garners the support of relevant stakeholders
- Outcome 4: Students and judges gain skills and insights into themselves and their abilities as well as practical legal issues.

The project involves activities including a mentorship programme that encompasses court visits, job shadowing and site visits, and internships at the Supreme Court of Appeal for students; women judges are provided with student interns and make use of their research skills; and judges and students participate in workshops and seminars.[14]

The main purpose of the evaluation is to assist the DGRU and the embassy in assessing the results (impact and outcome-level) of the project on its target group and to provide input to help the parties improve the project design for a possible new support period.[15] The evaluation was to assess achievements, identify strengths and weaknesses, and respond to evaluation questions linked to each of the four intended outcomes.

Both the report and executive summary keep to the page limit prescribed in the terms of reference. The summary is only one and a half pages, but still manages to cover all essential points. The report is well written and rich in detail. It presents a good discussion on findings, and conclusions and recommendations are well linked to findings.

The intended use and users of the report are stated, but the evaluation purpose and rationale, although stated in the executive summary, are not made explicit in the main body of the report. The programme, its components and execution are described well. Each component is described separately, with each section ending with findings. These sections include much detail about the programme, but few references to specific sources. This is a main shortcoming of the report.

The report provides a good presentation of how the programme intends to achieve its outcomes, both in the text and in a diagram and two tables in the annexes. It includes indicators and risks, but not assumptions. The programme theory is a bit vague but understandable, thanks to the evaluator's clear elaboration of the theory's components.

[12] Unknown author, 2020. End-term evaluation of Project RSA-3005, RAF-16/0046: Women Pioneers in The Judiciary - Supporting Female Judges and Female Law Students.
[13] Ibid, p. 6.
[14] DGRU, 2020, Terms of Reference: End-term evaluation of Project RSA-3005, RAF-16/0046: Women Pioneers in The Judiciary - Supporting Female Judges and Female Law Students, p. 1-2.
[15] Ibid, p.2.

The context could have been better and more concretely described. Bits and pieces of context appear throughout the report, but should have been presented in a more comprehensive way. While background, motivation for the project, etc. are provided, there is little information about external context.

The report provides a good description of the design of the evaluation, but does not explain why the evaluators selected this design; nor are the evaluation questions from the terms of reference repeated. Despite this, it is clear why the methods were selected and how they complemented each other. Triangulation is mentioned, and the approach is referred to as a multi-faceted approach.

Methods for data collection and analysis, and how interviewees were identified, are described. However, data collection tools (survey questions and interview questions) are not annexed to the report. The description of the survey contains a detailed account of how respondents were identified:

"By using the phone numbers on record at DGRU, email addresses for 37 students who had been mentees in 2017, 2018 and/or 2019 were found, and the link to an online survey sent to them with a request that they complete it. After two reminders had been sent, 32 responses had been received. This is an unusually high response rate for this type of survey and, in itself, suggests that the students valued the programme."[16]

The report comments on the response rate in the surveys and consequences are discussed. The evaluator considers the response rate surprisingly high and interprets this as an indicator of the good value of the project. There is a detailed account of the background of the survey respondents, including university affiliation, apartheid race classification and age. There is a list of persons interviewed, but no list of documents reviewed.

The methods seem well applied. A detailed account of data contributes to convincing the reader that data are valid and reliable. There is good referencing of interviews and survey results, survey results are well used, and interviews are referred to frequently. Although references to documents are missing, the data presented indicate that documents have been consulted. The main limitations mentioned refer to non-response in the survey and to limitations due to the COVID-19 pandemic. The consequences of these for the review are, however, not discussed, and the report does not mention ethical issues.

The report and terms of reference do not refer to evaluation criteria and cross-cutting issues by name, but still cover several of them.

Assessment of relevance is not required by the terms of reference but is nevertheless included in the discussion, and well handled. Although the term "effectiveness" is not used, the main task of the evaluation is to assess results, which is done and presented in detail in the report. Efficiency is also not included in the terms of reference but is briefly touched upon when the report compares costs for two programme components — a way of assessing efficiency that is neglected in many reports that are tasked with assessing efficiency.

The cross-cutting issue of "gender equality and women's rights" is a key motivation for the programme. It is well covered, although not specifically demanded in the terms of reference. The cross-cutting issue of "human rights" is also covered without having been requested. The following passage from the evaluation illustrates the complexity of this issue:

"Many interviewees said that while race was not explicitly named, they and the programme were not blind to it, and everyone understood that the programme would focus primarily on the previously disadvantaged. For example, the Norwegian Embassy said that the programme had 'nothing to do with race', but added that given that this was part of development assistance, the Embassy would not fund a programme benefiting people who were 'advantaged'".[17]

The report does a good job of responding to all questions in the terms of reference. Some pertinent questions, such as race and organisational complications, were added by the author and thoroughly discussed. Findings are clearly founded on evidence, conclusions are clear and based on findings, and recommendations are based on conclusions and on the discussions of findings.

[16] Unknown, 2020, p. 14.
[17] Ibid, p. 33.

Recommendations are actionable and targeted to users. They are clearly useful to guide a possible continuation of the project, both for the donor and for the implementing partners. The recommendations respond to the purpose of the review, and the author has added a strong focus on gender issues.

This is a very well-structured and readable report. It is unusual in the sense that is feels more like a narration than a presentation of data. Findings are well discussed, and the report avoids falling into the trap of presenting conclusions and recommendations here and there among findings. Shortcomings include lack of sources for documented evidence, lack of inclusion of data collection tools, and that evaluation questions are not presented. There also is no information about who conducted the review; no author or consultancy company is mentioned anywhere in the report.

The anonymous evaluator adds several important issues that are not mentioned in the terms of reference but seem highly relevant and useful. One, as mentioned, is the issue of race; another is the rather complicated organisational setup: Both are well discussed and provide relevant information. Also added are some stakeholder issues that seem important to implementation of the project, as the conclusion seem to be that it worked well despite the many different interests involved.

Despite the relatively small budget, this is obviously a relevant and strategic project, and the conclusion is that it worked well:

"The overall goal of the project is that women law graduates and judicial officers feel better equipped to fulfil their roles in the workplace. There can be no doubt that both the mentorship and internship programmes contributed in this way for many, if not most, of the students who participated. Several judges also confirmed how the interns had assisted them in their own work. The fact that the [Supreme Court of Appeal] and the High Courts would like to expand the internship programme attests to the fact that, overall, judges feel that the interns' work facilitates their own legal work. The main question relating to the goal is that the programme involved men alongside women in most parts of the project."[18]

[18] Ibid, p. 37.

Annex 7: Profiles of the assessment team

Ms Ingela Ternström, Team Leader

Ms Ternström holds a PhD in Economics with a focus on Environmental and Development Economics and a MSc Business and Economics, both from the Stockholm School of Economics. Ms Ternström brings eight years of postdoctoral research experience at the Royal Swedish Academy of Science in the field of natural resource management, as well as field experience of development cooperation and research. Over the past decade, Ms Ternström has worked as a full-time senior consultant with Ternstrom Consulting AB, where her main role has been as quality assurer and methodology expert. In addition, Ms Ternström has been team leader, project manager and team member on a range of evaluations, including several evaluations for Norad and the Evaluation Department in Norad.

Ms Ternström's education and experience bring an in-depth understanding of methodological and practical issues relating to evaluation quality. These include selection of approaches and data collection methods, experience of practically applying them to evaluation work, quality assuring evaluation teams and reports, and assessing the quality of evaluation reports. Ms Ternström understands the complexities involved in combining methodological rigor with practical challenges involved in evaluating development cooperation. Ms Ternström has experience of working with and managing diverse teams, large volumes of data, quality issues, and complex evaluation processes.

Mr Stefan Dahlgren, Team Member

Mr Dahlgren, BA in Sociology, has 15 years of experience as an evaluation manager, four of these as Director of the evaluation department at the Swedish International Cooperation Agency (Sida). Mr Dahlgren has gained extensive experience and understanding of development evaluation, where quality assessment of the reports and the evaluation process is essential. He has a broad background established in different roles in development cooperation and evaluation as well as in a number of themes and contexts where evaluations are carried out.

Mr Dahlgren's professional background includes roles as external evaluation consultant, programme officer for handling of implementation and follow-up of projects and programmes in various countries, and head of development cooperation at different embassies in Africa and Asia with responsibility to discuss policies and strategies with partner countries. In both the latter roles, Mr Dahlgren was thus recipient and user of evaluations. Mr Dahlgren's assignments as a consultant include the role of overall team leader for six assessment teams and leader of the quality assurance group of a programme to assess 30 non-governmental organisations (NGOs) for Sida.

Before joining Sida, Mr Dahlgren worked for ten years as a researcher (team leader and head of department) at the Swedish National Housing and Building Research Institute. Mr Dahlgren's research background provides a solid methodological knowledge in applied social research.

Mr Jock Baker, Team Member

Mr Baker holds a MSc in Economics from the London School of Economics. An independent consultant with more than ten years' experience in senior monitoring and evaluation/Quality Assurance roles with CARE USA and CARE International, as well as 20 years of field experience, including 13 years in humanitarian operations with UNHCR, UNOPS/UNDP and OCHA, Mr Baker brings in-depth experience of humanitarian issues from both operational and Quality Assurance roles.

As a consultant, Mr Baker has been involved in more than 50 evaluations and was team leader for most of them. Mr Baker has also been engaged in developing the methodological toolbox used by stakeholders in humanitarian response, including six years on the Board of the Assessment Capabilities Project (ACAPS). Mr Baker has also published on a series of humanitarian issues such as value for money in the humanitarian sector, strong and weak points of the core humanitarian standard and humanitarian capacity building in collaboration as well as a proposal for a methodology to cost interagency humanitarian response plans. Mr Baker's combination of operational experience and experience of commissioning, managing, implementing and

Quality Assuring evaluations, as well as publishing on methodological issues relevant to the sector, have given him a thorough understanding end expertise that is highly relevant for this assignment.

Ms Eva Lithman, Team Member

Ms Lithman, with an MA from Stockholm University, is an experienced and versatile evaluation professional with extensive experience of evaluation and performance auditing as project manager, quality reviewer and manager. Ms Lithman has operational experience from bilateral, multilateral and civil society organisations and has lived in South America, Central Asia and North America. She is fluent in English, French and Spanish.

Ms Lithman has been Audit Director at the Swedish National Audit Office, Director of Sida's Evaluation department and Chair, OECD DAC Evaluation Network. After retiring from Sida in 2012, Ms Lithman worked extensively as an evaluation quality reviewer covering a broad set of themes in addition to being a member of the Expert Group for Aid Studies, (EBA) commissioned to analyse and evaluate Swedish international development cooperation. In recent years, Ms Lithman has been increasingly engaged in evaluation of action related to climate change, currently as chair of the newly created Adaptation Fund Technical Evaluation Reference Group.

Mr Abid Rehman, Team Member

Mr Rehman is a Pakistani national with a Master's in Political Science. He is a monitoring, evaluation and reporting professional with significant programme development experience. Over the past 12 years. Mr Rehman has successfully designed and managed monitoring and evaluation learning-focused teams for large-scale governance, peacebuilding, conflict resolution and stabilisation programmes in politically dynamic environments including Pakistan, Afghanistan and South Sudan, including 18 months as Monitoring and Evaluation Manager of USAID's USD 110-million flagship peacebuilding, stabilisation, and conflict mitigation programme in South Sudan.

Mr Rehman has applied, managed and quality assured a range of data collection systems and methodologies including surveys, capacity building for distance management of monitoring, iterative outcome mapping and quality assurance of data management systems. With an in-depth understanding of monitoring and evaluation systems and processes, Mr Rehman is skilled in developing, refining and integrating monitoring and evaluation tools that support active learning and adaptation of humanitarian response programming implemented in non-permissive and highly volatile security environments.

Mr Abhijit Bhattacharjee, Quality Assurer

With an MSc in Agricultural Economics and Statistics and more than 36 years of development and humanitarian sector work, Mr Bhattacharjee has extensive professional experience in undertaking reviews and evaluations in various settings for the United Kingdom Department for International Development, the European Union/ECHO, UN agencies (UNDP, UNHCR, OCHA, UNICEF, World Food Programme), NGOs, consortiums, and networks. He is currently providing team leadership for Sida's humanitarian partnership evaluation. He recently (2018-19) a major evaluation for the EU combining Afghanistan country portfolio evaluation (2014-19) and ECHO's global partnership with Norwegian Refugee Council. Besides leading and managing evaluations, reviews and systematic studies, Mr Bhattacharjee has undertaken quality assurance functions for several major evaluations, including UNDP's evaluation of the Comprehensive Disaster Management Programme (2016) UNDP's evaluation of the country programme in Pakistan (2016-17); and UNHCR's evaluation of the country programme in South Sudan and the Democratic Republic of Congo (2019).

Mr Bhattacharjee brings extensive evaluation skills and experience combining qualitative and quantitative methodologies in social science research to ensure high standards of analysis of evidence. He further brings advanced knowledge and experience in using various international standards and methodologies, among them OECD DAC criteria, Core Humanitarian Standard, Red Cross Red Crescent/NGO code, UNEG standards and evaluation ethics, ALNAP Quality proforma, and Sphere, BOND Evidence Principles.