Norad

EVALUATION DEPARTMENT

Report 3 / 2021

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)

This report is the product of the authors, and responsibility for the accuracy of data included in this report rests with the authors alone. The findings, interpretations, and conclusions presented in this report do not necessarily reflect the views of the Evaluation Department.

October 2021

Table of contents

Preface 4
Acknowledgments 5
Executive summary 6

1 Background and purpose 8

2 Approach and methodology 10
2.1 Quality assessment tool 10
2.2 Assessment and analysis process 11
2.3 Limitations and ethical issues 12

3 Overview of the quality of reports and terms of references 13

4 Quality in key focus areas 19
4.1 The foundation for the evaluation 20
4.2 Evaluation design and methods 24
4.3 Cross-cutting issues, ethical considerations and quality assurance 27

5 Conclusions 30

References 32
Annex 1: Terms of references 33
Annex 2: Descriptive statistics 39
List of annexes 41
Acronyms and abbreviations 42
Related reports 43

List of figures

Figure 1: Distribution of scores for evaluation reports 14
Figure 2: Average scores for report quality criteria 16
Figure 3: Quality areas for evaluation reports 19
Figure 4: Foundation, implementation and results of the evaluation 20

List of tables

Table 1: Average and standard deviation, report quality criteria 39
Table 2: Average and standard deviation, terms of reference quality criteria 40

Preface

Most evaluations commissioned by the Norwegian Aid Administration are initiated by the units responsible for grant management in the development aid administration. These evaluations - commonly called decentralised evaluation or reviews - are intended to form a key part of the evidence base for documenting results of Norwegian development cooperation. An evaluation conducted by the Evaluation Department in 2017 found the quality of the decentralised evaluations to be low and questioned the extent to which they provided credible information about results. A study published in 2020 found that the quality had not improved and this study reveals that the overall quality of decentralised evaluations is still low.

For decentralised evaluations to fulfil its intention, we encourage the Ministry of Foreign Affairs and Norad to strengthen its efforts to improve the quality of these evaluations. We believe our study can feed into this work. The study was carried out by the Swedish consultancy company Ternstrom Consulting AB and we thank the team for a job well done.

Oslo, 10th September 2021

Siv Lillestøl
Acting Director, Evaluation Department

Acknowledgements

This report summarises the results of the quality assessment of 27 decentralised evaluations commissioned by the Norwegian aid administration and published in 2019 and 2020. The team is grateful for this opportunity to gain insights into a large variety of programmes. We are impressed by the richness of information presented in the reports and by the achievements of the evaluated interventions and acknowledge the hard work of commissioners, evaluators and programme staff.

This assignment was implemented by Ternstrom Consulting AB. The team consisted of Ms Ingela Ternström (Team Leader), Mr Jock Baker, Mr Stefan Dahlgren, Ms Eva Lithman and Mr Abid Rehman.

Quality assurance was provided by Mr Abhijit Bhattacharjee. All team members participated in the quality assessment of reports. The report was prepared by the Team Leader with assistance from other team members and following invaluable discussions among the team.

Täby, August 2021

Ingela Ternström

Executive Summary

This study assesses the quality of decentralised evaluations of Norwegian development cooperation. Decentralised evaluations are an important source of information about the results of development projects and programmes and are used for learning and accountability purposes. The credibility and utility of decentralised evaluations are therefore important. In this study, the quality of decentralised evaluations that were commissioned by the Norwegian aid administration and completed in 2019–2020 is assessed. It is a follow-up to the study commissioned by the Evaluation Department in Norad and published in 2020, which found that the quality of decentralised evaluations published in 2018–2019 fell short, especially in terms methodology, transparency regarding quality of data and attention to cross-cutting issues and ethical considerations. The same shortcomings were identified by an evaluation published by the Evaluation Department in Norad in 2017.

The purpose of this and last year's study is to assess quality and provide quality assurance units in the Ministry of Foreign Affairs and Norad with information about strengths and weaknesses of these evaluations, which in turn can be used to improve the quality of evaluations. In addition, this study considers the credibility of information presented in decentralised evaluations. Finally, the publication of an annual report may make quality a greater focus of commissioners' and evaluators' attention.

The report summarises the results of the quality assessment of 27 decentralised evaluations and 24 terms of references, commissioned by the Norwegian aid administration and published in 2019 and 2020.

The quality assessment is limited to information presented in written evaluation reports and terms of references. Other aspects of the evaluation process are not included in the study. A standardised rating manual was used to assess quality. The quality criteria set out in the rating manual are largely based on the OECD Development Assistance Committee's quality standards for evaluations and evaluation criteria, as well as on cross-cutting themes defined by the Ministry of Foreign Affairs. The rating manual was developed by the study published by the Evaluation Department in Norad in 2017 and revised by the Evaluation Department in Norad in August 2020 to improve rating guidance and to better reflect the Evaluation Department's priorities. The quality assessments were made by a team of five highly experienced evaluation professionals. Extensive measures were taken to ensure consistent application of the rating manual, including pilot scoring and team discussions of three reports and double scoring of additional evaluation reports and terms of references.

The revision of the rating manual makes exact comparisons with the previous quality assessments difficult. However, the results do not indicate an improvement in quality from what was found in 2020. The overall quality of decentralised evaluations is still below par. Nearly 40 percent of the reports had poor or inadequate quality on a majority of the quality criteria; nearly 20 percent of the reports had poor or inadequate quality on more than 90 percent of the quality criteria. The quality of terms of references was more uniform: All of them had satisfactory quality on between 25 and 75 percent of the quality criteria.

The results show that quality was highest on presenting recommendations and conclusions and on responding to evaluation questions. The quality was lower on linking conclusions to findings and findings to evidence, and was even lower on providing sources for the evidence that was presented. Evaluation design and methods for collecting and analysing data were often very briefly described and rarely discussed or justified, and the quality of data was most often not discussed. Ethical issues and cross-cutting issues were consistently neglected.

The study thus indicates that while the evaluation reports do a good job of responding to the purpose of the evaluations, they less frequently explain how they arrived at these responses. This limits the credibility of the evaluations and makes learning from evaluations risky. The lack of contextual information and details about programme theory in many reports makes it difficult to interpret the reports unless you are already familiar with the programme. This limits the scope for learning and the extent to which the evaluation is useful for outsiders and future staff. The scope for learning is also limited by the fact that most evaluation reports were not publicly available.

This study was conducted by Ternstrom Consulting AB on behalf of the Evaluation Department in Norad.

Background and purpose

This report presents the findings of an independent quality assessment of 27 decentralised evaluations and reviews commissioned by the Norwegian aid administration[1] and finalised during 2019–2020. The report is a follow-up to the quality assessment of 55 decentralised evaluations implemented in 2018–2019 that was carried out by the same team in 2020 and presented in Norad (2020).[2]

Evaluations have multiple functions in aid administration, including ensuring accountability to donors, partners and beneficiaries; assessing achievements; and making recommendations regarding programme management and future funding. Evaluations thus have the potential to affect funding and implementation in a large number of interventions, with extensive impact on implementing organisations and target populations worldwide. To ensure that correct decisions are made, it is crucial that evaluation reports are credible and useful. Decision makers should be able to use the evaluation with a high degree of confidence. Thus, the data, evidence and findings that form the basis for conclusions and recommendations need to be of high quality and present a reliable and unbiased picture of reality. This requires that the methodology used to select, collect and analyse information be appropriate to the task at hand.

Achieving adequate quality of decentralised evaluations is challenging for the Norwegian aid administration, just as it is for many other agencies.[3] This was illustrated in Norad (2020), which found the quality of evaluations to be particularly low in terms of methodology and transparency, with methods poorly described and applied and with little transparency about shortcomings and limitations of the analysis. It further found the link from evidence to findings and conclusions was not sufficiently clear, ethical considerations were rarely mentioned, and efficiency was often poorly assessed. Norad (2017),[4] which assessed the quality of decentralised evaluations implemented in 2014, identified very similar shortcomings.

The present assessment was commissioned by the Evaluation Department in Norad as part of an effort to improve the quality of reviews and decentralised evaluations commissioned by the Norwegian aid administration by providing an annual diagnostic of the quality of published reviews and decentralised evaluations, thus contributing to both learning and accountability. Three objectives are identified in the Terms of Reference (presented in Annex 1):

[1] The aid administration here refers to the Ministry of Foreign Affairs, Norwegian embassies and Norad.

[2] Norad (2020). Quality Assessment of Decentralised Evaluations in Norwegian Development Cooperation (2018–2019). Evaluation Department Report 6/2020, Norad.

[3] OECD DAC (2016). Evaluation Systems in Development Co-operation: 2016 Review. OECD Publishing, Paris.

[4] Norad (2017). The Quality of Reviews and Decentralised Evaluations in Norwegian Development Cooperation. Evaluation Department Report 1/2017, Norad.

— Assess the quality of reviews and decentralised evaluations of Norwegian development cooperation (using a pre-defined format).

— Identify strengths and weaknesses of reviews and decentralised evaluations.

— Summarise findings from the reviews and decentralised evaluations, taking into consideration their credibility and based on the assessed quality.[5]

While the main intended users of this report are the quality assurance units of the Ministry of Foreign Affairs (MFA) and Norad, the information may also be useful for other parts of the Norwegian aid administration, to make quality a greater focus of commissioners' and evaluators' attention.

The object of the assignment is decentralised evaluations commissioned by the Norwegian aid administration. The scope is limited by year of publication (2019 or 2020), source of funding (Norwegian development cooperation), and commissioner (Norad, the MFA or Norwegian embassies[6]). Both mid-term and final reviews and evaluations and internal and external evaluations were included.[7] It should be noted that the scope includes evaluation reports and terms of references only; tenders, inception reports and other aspects of the evaluation process are not included.

The requirement to conduct evaluations follows from the Regulations for Financial Management in the Government Administration (Økonomiregleverket). The Ministry of Finance's definition of evaluation encompass reviews, and the term "decentralised evaluation" is used henceforth for both reviews and decentralised evaluations.[8] Occasionally, the terms "evaluation" and "evaluation report" are used for ease of reading.

Chapter 2 of this report describes the approach and methodology; complementary information is presented in Annex 3. Chapter 3 provides an overview of the quality of assessed reports and terms of references and Chapter 4 looks closer at selected areas where the need for improvement is especially high. Chapter 5 summarises the team's findings and conclusions drawn. Additional detail is provided in the annexes: Annex 4 presents a list of included evaluations; Annexes 2 and 5 present data and Annex 6 presents summaries of the three reports with highest scores. Annex 7 presents the assessment team. Annexes 3 to 7 may be found in a separate document together with this report at norad.no/evaluation.

[5] Due to the low quality of the assessed evaluations, the third objective was revised and instead, the three reports with highest scores are summarised to illustrate best practice in terms of quality. This change was made in agreement with the Evaluation Department in Norad.

[6] Evaluations commissioned on behalf of the Ministry of Climate and Environment were also included.

[7] These criteria were agreed upon during the inception phase.

[8] These exclude evaluations commissioned by the Evaluation Department, which are referred to as independent evaluations.

Approach and methodology

The assignment was implemented by a team of highly experienced evaluation professionals. The team members' combined areas of competence have contributed to a high level of understanding of the many contexts, methods and thematic areas of the decentralised evaluations that were assessed.[9]

The approach consists of a standardised assessment of the quality of reports and terms of references of decentralised evaluations[10] using a strict, predefined tool; processes to ensure that the application of the tool was as consistent as possible across raters (quality assessors) and evaluation reports and over time to reduce bias; and a well-functioning structure for retrieving and storing reports and collating and analysing data.[11] A document review of other assessments of evaluation quality, including for example peer reviews by the Organisation for Economic Co-operation and Development (OECD) and other donors' quality assessments provided input to the development of the approach, methodology and tools.[12] An inception report was presented to and approved by the Evaluation Department.[13]

2.1 Quality assessment tool

The quality assessment tool consists of a rating manual for reports and terms of references that is based on the OECD Development Assistance Committee (DAC) quality standards for evaluating development assistance.[14]

The rating manual contains separate quality criteria for reports and terms of references. It was originally developed by the authors of Norad (2017) and used in Norad (2020). The rating manual for evaluation reports was revised in August 2020 by the Evaluation Department in consultation with the assessment team. The main changes from the 2020 assessment were the removal of overlapping quality criteria and of criteria that were irrelevant; the addition of a few quality criteria to better reflect the priorities of the Evaluation Department; and improvement in the scoring guidance to enable consistent rating. In addition, it was decided that to encourage learning, scoring protocols will be shared with evaluation commissioners and the quality assurance units at Norad and MFA.

The rating manual uses a four-point scale, with scores 1 (poor quality) and 2 (less than adequate quality) indicating less than satisfactory quality and scores 3 (adequate quality) and 4 (good quality) connoting satisfactory quality. The rating manual presents a qualifying statement and detailed scoring guidance

[9] Brief bios of the team members can be found in Annex 7.

[10] As noted in Chapter 1, the term "decentralised evaluations" is used in the remainder of this report to refer to both reviews and decentralised evaluations.

[11] The approach is summarised in a brief evaluation matrix in Annex 3.

[12] See, for example, OECD DAC (2016), OECD (2019a), Department of Foreign Affairs and Trade: Australian Government (2018), Cooney, Rojas, Arsenault and Babcock (2015).

[13] The inception report includes descriptions of approach, methods, evaluation matrix, limitations, ethical considerations as well as a description of tools and copies of formats used.

[14] OECD (2010), Quality Standards for Development Evaluation, DAC Guidelines and Reference Series, and OECD (2019b), Development Assistance Committee: Better Criteria for Better Evaluation.

for each quality criterion.[15] The rating manual was translated into scoring protocol formats, with space added for administrative data, key findings from the report and the rater's general comments. These are presented in Annex 3. During the piloting and calibration process, the rating manual was thoroughly discussed and, where needed, complemented with clarifying comments in the scoring protocol formats.

As only one relevant report had been published on Norad's webpage in 2020, the retrieval of reports and their corresponding terms of references was done almost exclusively through direct contact with commissioners. An email was sent to 81 potential evaluation commissioners, requesting them to submit decentralised evaluation reports and terms of references to the team. A copy of the request is available in Annex 3.[16] A total of 27 evaluations that fit the scope of the assignment (see Chapter 1) were received and quality assessed. Terms of references were available for 24 of these. Annex 4 presents the list of included evaluations, including information about commissioner, implementer, thematic area and country.

2.2 Assessment and analysis process

The team devoted a considerable amount of time to ensuring consistency in rating across raters in preparation for the first round of quality assessments (in 2020). Renewed piloting and calibration were made in 2021, starting with a team meeting to discuss the revised quality assessment tool. This was followed by pilot rating and team discussions to arrive at consensus scores for three pilot evaluations. In addition, two evaluations were double scored (scored by two raters). All meetings were conducted remotely via Zoom, with the Evaluation Department participating in the first of the three pilot report meetings.[17]

All team members participated in the quality assessment process, making it possible to draw upon a rich pool of experience from various geographic and thematic areas of the aid sector. To the extent possible, team members' experience and competencies were matched to the theme and context of the evaluated interventions. To reduce potential sources of bias, reports authored by the same firm were distributed to a range of different raters, and raters were asked to declare any prior relationships with report authors.[18]

Rating protocols were prepared for each report and terms of reference, uploaded to an Excel database, and reviewed by the team leader and the team's quality assurer. Where there were inconsistencies in justifying comments and scores, these were commented on and returned to the raters for revision or explanation. As a final quality assurance mechanism, the team leader cross-checked the scoring of five reports.[19]

[15] The original quality assessment tool did not include scoring guidance for terms of reference quality. The assessment team developed this guidance in August 2020.

[16] The initial send list included 43 embassies and foreign missions, 22 sections or departments at the MFA and 16 sections or departments at Norad. Both Norad and the MFA sent a reminder following the intial request to an extended and revised list of recipients. Replies were received from 36 of these, either by sending reports or by stating that they have not commissioned any reports.

[17] Pilot and double scoring reports were selected by the Team Leader based on length and topic to enable a focus on quality criteria and a first-glance assessment of quality to ensure the inclusion of both high- and low-quality reports.

[18] Potential bias due to prior relationships with evaluation authors was avoided by asking raters to notify the team leader of such cases and cross-check rater CVs.

[19] These five reports were selected to include at least one report per scorer and guided by the quality assurer's review.

Quantitative information was analysed using Excel features to develop descriptive statistics of the quality of reports and terms of references and to compare scores *within* and *between* quality areas.[20] Qualitative information was analysed through a review of justifying comments for each quality criterion to assess the reliability of scores and collect examples for the findings chapter of this report. Evaluation reports and terms of references were consulted for further examples and background information, and the three evaluation reports with the highest scores were summarised to illustrate examples of good practise (see Annex 6). Finally, discussions within the assessment team have been an important part of the analysis process.

2.3 Limitations and ethical issues

The submission of evaluation reports to the team was voluntary. The team has not been able to obtain information about the total number of decentralised evaluations that were completed during the period of review and does not know how the submitted evaluations were selected. Hence, it is not possible to assess whether the quality of the submitted evaluations is representative of all decentralised evaluations or if, for example, reports of poorer quality were not included in the sample. The quality assessments are based on a rigorous assessment method with pre-defined criteria; while some subjectivity may remain due to rater bias, substantial effort has been taken to minimise this through the quality control and assurance process described above. The team assesses that the validity and reliability of the assessment data are sufficient for drawing conclusions about the quality of evaluations. Ethical issues and safeguards, and further details about limitations and challenges, are described in Annex 3.

The aim of this assignment reflects an effort to improve the quality of reviews and decentralised evaluations. This report provides information about the quality of decentralised evaluations and to some extent about the consequences of their strengths and weaknesses. Whether this contributes to learning and to improved quality of decentralised evaluations is outside the control of the authors.

[20] Comparison of scores across raters to identify remaining bias was not feasible due to the low number of reports per rater.

Overview of the quality of reports and terms of references

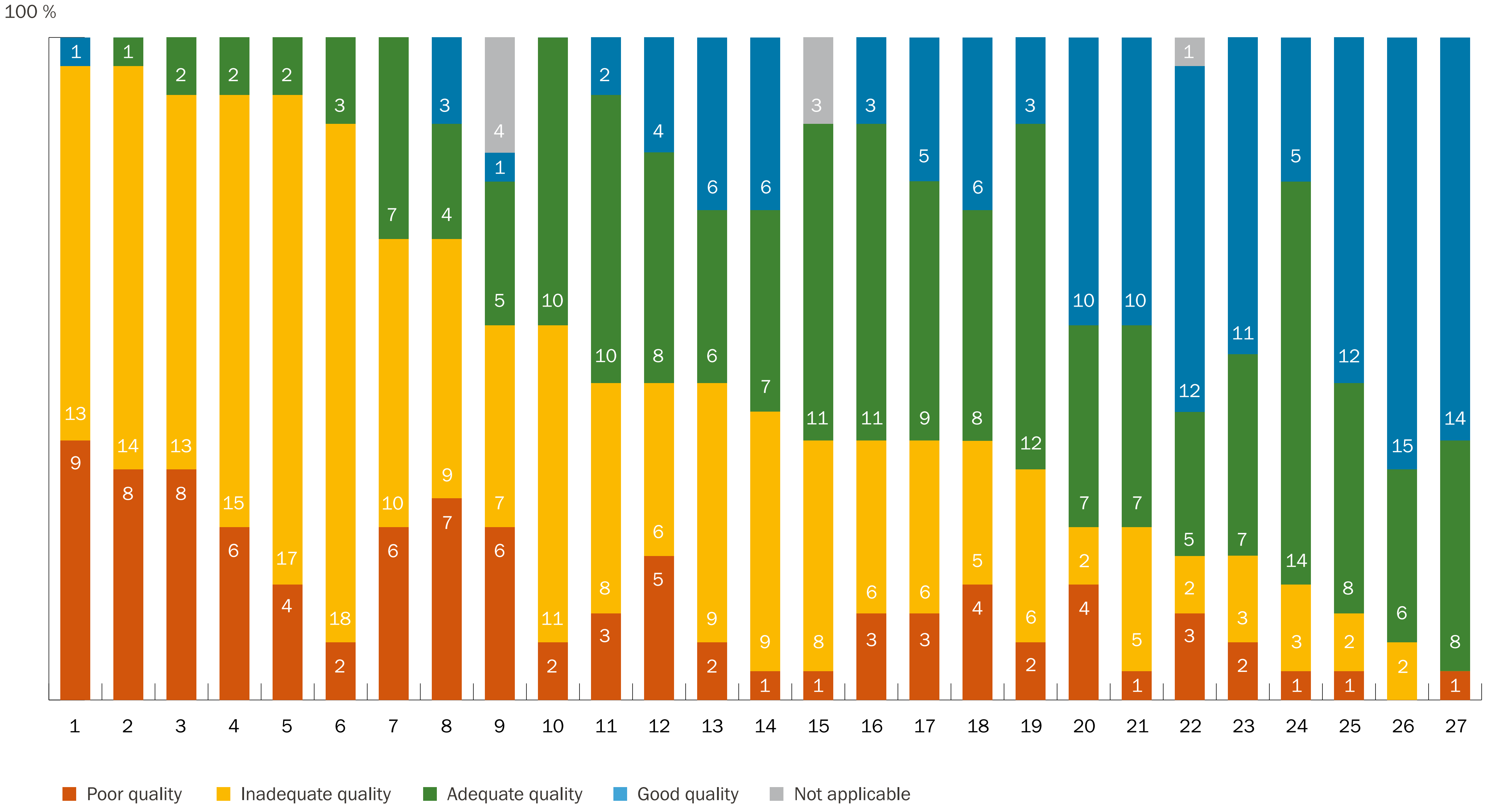
This chapter provides a brief overview of the results of the quality assessment of reports and their corresponding terms of reference. Annex 2 presents descriptive statistics and Annex 5 presents data and diagrams. The information here builds on scores and justifying comments made by the raters. A total of 27 evaluations are included; for 24 of these, the team had access to the terms of references.[21] As the quality assessment tool has been revised, results on single quality criterion cannot be compared across years.

Finding 1: More than a third of the reports had poor or inadequate quality on a majority of the quality criteria.

Figure 1 (next page) shows the quality of each evaluation report. The colours and numbers on the columns indicate the number of quality criteria with poor quality (score 1, orange), inadequate quality (score 2, yellow), adequate (score 3, green) and good quality (score 4, blue). The reports are sorted by the sum of scores 1 and 2 so that the reports with lowest quality are to the left in the diagram. Ten reports (37 percent) have poor or inadequate quality (score 1 or 2) on more than half e quality criteria; eight reports (30 percent) have poor or inadequate quality on more than two-thirds of the quality criteria; and five reports score 1 or 2 on more than 90 percent of the quality criteria. On the other hand, 17 reports (63 percent) score 3 or 4 (adequate or good quality) on at least half of the quality criteria. Eight reports (30 percent) score 3 or 4 on more than two-thirds of the quality criteria, and two reports score 3 or 4 on over 90 percent of the quality criteria.

[21] The team put considerable efforts into requesting terms of references for reports that were submitted without them. As a result, a larger share of reports were complete, with their terms of references, than was the case last year.

Figure 1. Distribution of scores for evaluation reports

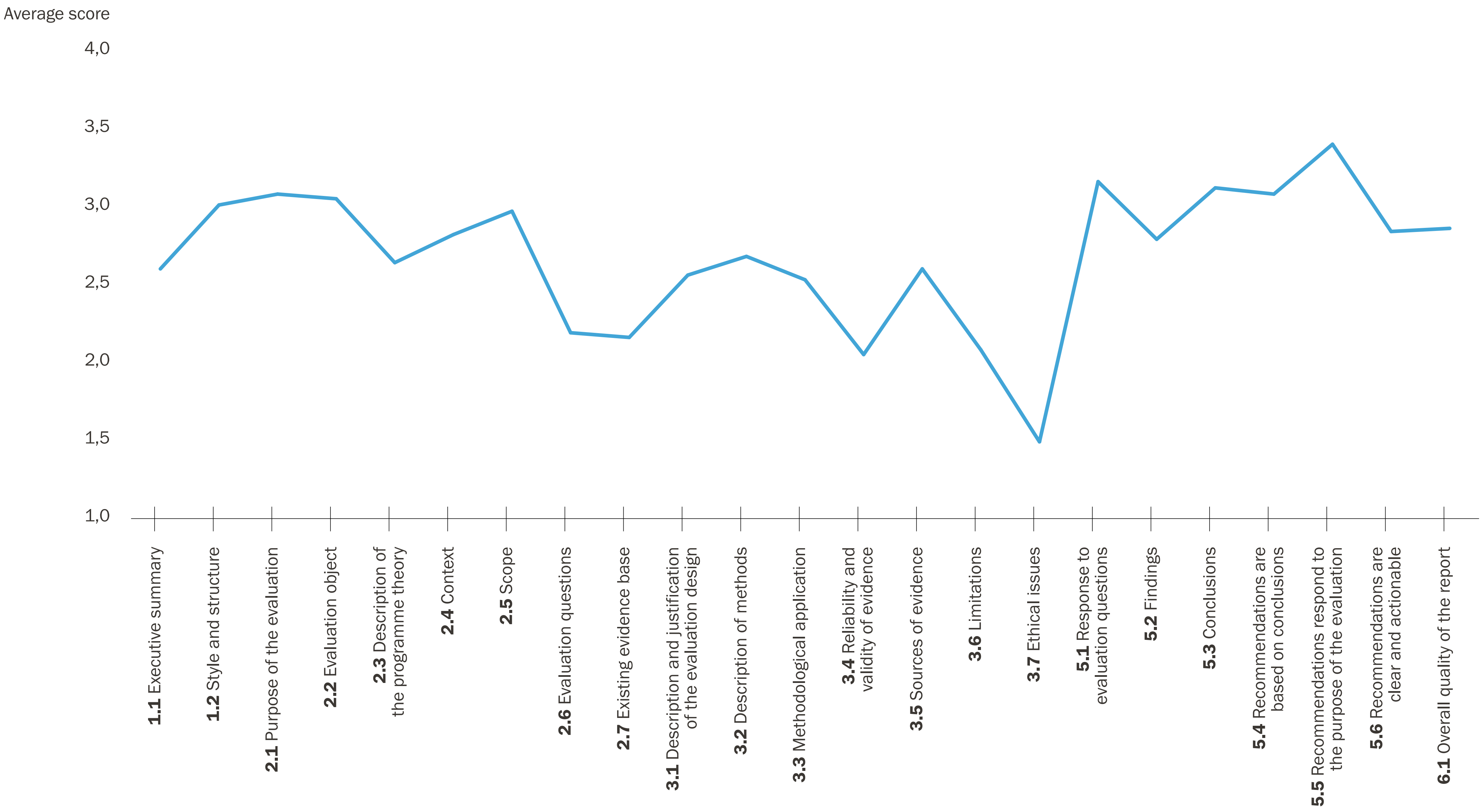


Note: Each column in the figure illustrates the distribution of scores for one report. The orange part of the column shows the number of quality criteria with poor quality (score 1), the yellow part shows the number of quality criteria with inadequate quality (score 2), the green part shows the number of quality criteria with adequate quality (score 3) and the blue part shows the number of quality criteria with good quality (score 4). In a few cases a quality criterium was not applicable (illustrated in grey), for example if the terms of reference did not ask for recommendations. The reports are sorted by the share of quality criteria with score 1 or 2, so that the report with the largest share of low scores is illustrated by the left-most column.

Figure 2 (next page) shows the average score for the different report quality criteria.[22] Due to the changes in the scoring tool in August 2020, a detailed comparison with last year's quality assessment is not possible but the overall pattern is the same. The average score is higher for quality criteria relating to presentation of the evaluation object and the purpose and scope of the evaluation and for criteria relating to responding to evaluation questions and presenting findings, conclusions and recommendations. The average score is lower for quality criteria relating to how the evaluation was implemented (design, methods and sources). Description of ethical issues has the lowest quality, followed by discussions of limitations and reliability and validity.

[22] The average score for a quality criteria was calculated as the sum of all reports' scores on the quality criteria, divided by the number of reports.

Figure 2. Average scores for report quality criteria



Note: The line illustrates the average score on each quality criteria. The average on a specific criterion is calculated as the total of the 27 reports' scores on this criterion and divided by 27. The scores range from 1 to 4. Score 1 indicates poor quality, score 2 inadequate quality, score 3 adequate quality and score 4 indicates good quality.

Some overall observations are presented below. A selection of key issues is discussed in more detail in Chapter 4.

Finding 2: Evaluation report authors are good at responding to evaluation questions and presenting conclusions and recommendations, but not as good at explaining how they arrived at these.

The three quality criteria with the highest average scores in this assessment are relating to responding to evaluation questions, presenting conclusions, and presenting recommendations that respond to the purpose of the evaluation (criteria 5.1, 5.3 and 5.6). These three quality criteria have the largest number of reports that score 4 and the highest average scores, and they are the only quality criteria that have average scores above 3. The evaluation reports are also good at linking recommendations to conclusions, but working backwards from recommendations towards evidence, the quality decreases.

For example, ten reports (37 percent) present conclusions to evaluation questions that flow clearly and logically from the analysis of findings. In seven reports (26 percent), conclusions follow from findings, but with some gaps. In the remaining ten reports (37 percent), conclusions can only be partially traced back to analysis of findings.

The reports are poorer at linking findings to evidence. Only five reports presented findings that were clearly founded on evidence; over half (14) of the reports presented findings that were only partially founded on evidence, with many gaps. The quality is also low on providing sources for the evidence that is the foundation for findings, conclusions and recommendations. Sources of evidence are poorly referenced in 60 percent of the reports; only two reports had consistent and clear referencing to the sources of evidence.

Finding 3: There is a lack of transparency regarding sources of information and quality of data.

Only five of the reports described how sources of information were selected (e.g. selection of interviewees, sampling of survey respondents, why specific locations were visited). In several evaluations, the terms of reference suggest or prescribe which stakeholders should be interviewed; few reports comment on this or discuss the risk for bias in letting the commissioner control the selection of sources.

Reliability and validity of evidence – that is, to what extent the data used for arriving at findings, conclusions and recommendations give an accurate and unbiased picture of reality – were poorly discussed, or not discussed at all, in 70 percent of the reports. Only three reports provided a thorough discussion of reliability and validity.[23] Limitations are another area that is poorly discussed in the reports. Most frequently, the reports referred to limitations to the evaluation in terms of time or logistical challenges. Limitations arising from the way the evaluation was implemented were rarely mentioned.

[23] The OECD (2002) glossary defines validity as the "extent to which the data collection strategies and instruments measure what they purport to measure" and reliability as "consistency or dependability of data and evaluation judgements, with reference to the quality of the instruments, procedures and analyses used to collect and interpret evaluation data".

Finding 4: More than two-thirds of the terms of references had adequate or good quality on a majority of the quality criteria.

A majority of the terms of references (71 percent), had satisfactory quality (score 3 or 4) on at least half of the quality criteria. There were fewer terms of references than reports with very low or very high quality. All terms of references had good or adequate quality on at least 25 percent of the quality criteria; none had poor or inadequate quality on more than 25 percent of the criteria.

Finding 5: Evaluation commissioners are better at expressing what they want the evaluation to deliver than at providing information about aspects that the evaluation team needs to take into consideration.

The terms of references score high on quality criteria that refer to the description of rationale and purpose, objectives, evaluation criteria (e.g. relevance, effectiveness, sustainability, etc.), and evaluation questions and to what is wanted from evaluators in terms of deliverables. The quality is lower on criteria that refer to background information that the evaluators need for planning the evaluation such as context, previous evaluations of the intervention and expected limitations to the evaluation.

Cross-cutting issues, which must be considered in all aspects of Norwegian development aid, are often neglected in the terms of references. Gender issues are mentioned in 58 percent of the terms of references; human rights issues and environmental issues in 50 percent; and anti-corruption in only 37 percent. The average scores on cross-cutting issues range from 1.6 to 2.0. The two criteria with the lowest average scores are ethical issues and quality assurance (the average score is 1.42 on both); 18 of the terms of references (75 percent) do not at all mention quality assurance; 19 of them (79 percent) do not mention ethical issues.

Quality in key focus areas

This chapter presents findings on a selection of key issues and discusses the consequences of the observed strengths and weaknesses. The quality criteria for evaluation reports are grouped in five quality areas (Figure 3). While quality areas 1 and 4 refer to the overall presentation and to the evaluation criteria covered, quality areas 2, 3 and 5 refer to different aspects of the evaluation process.[24] Quality area 2 includes description of evaluation purpose, questions and object as well as context and other things that should be considered by the evaluators. These criteria can be seen as the foundation of an evaluation: They set the scene and state the conditions for the evaluation. Quality area 3 refers to how the evaluation was implemented and how evidence was translated into findings, taking the foundations into consideration. Quality area 5 refers to the results of the evaluation process, such as response to evaluation questions, conclusions and recommendations. The evaluation process can be seen as a pyramid, with the bottom of the pyramid as the foundation, represented in quality area 2; the mid-section of the pyramid is the implementation of the evaluation, captured in quality area 3; and the results of the evaluation are shown as the top of the pyramid. As presented in Chapter 3, the quality assessment shows that the reports score high on the top part of the pyramid but that the base and mid-level of the evaluation process are of lower quality.

Figure 3. Quality areas for evaluation reports

QUALITY AREA GROUPINGS

Quality area 1: Overall presentation

Quality area 2: Presentation of the purpose and object of evaluation

Quality area 3: Presentation of methodology

Quality area 4: Evaluation criteria

Quality area 5: Findings, conclusions and recommendations

[24] For more detail, see Annex 3.4.

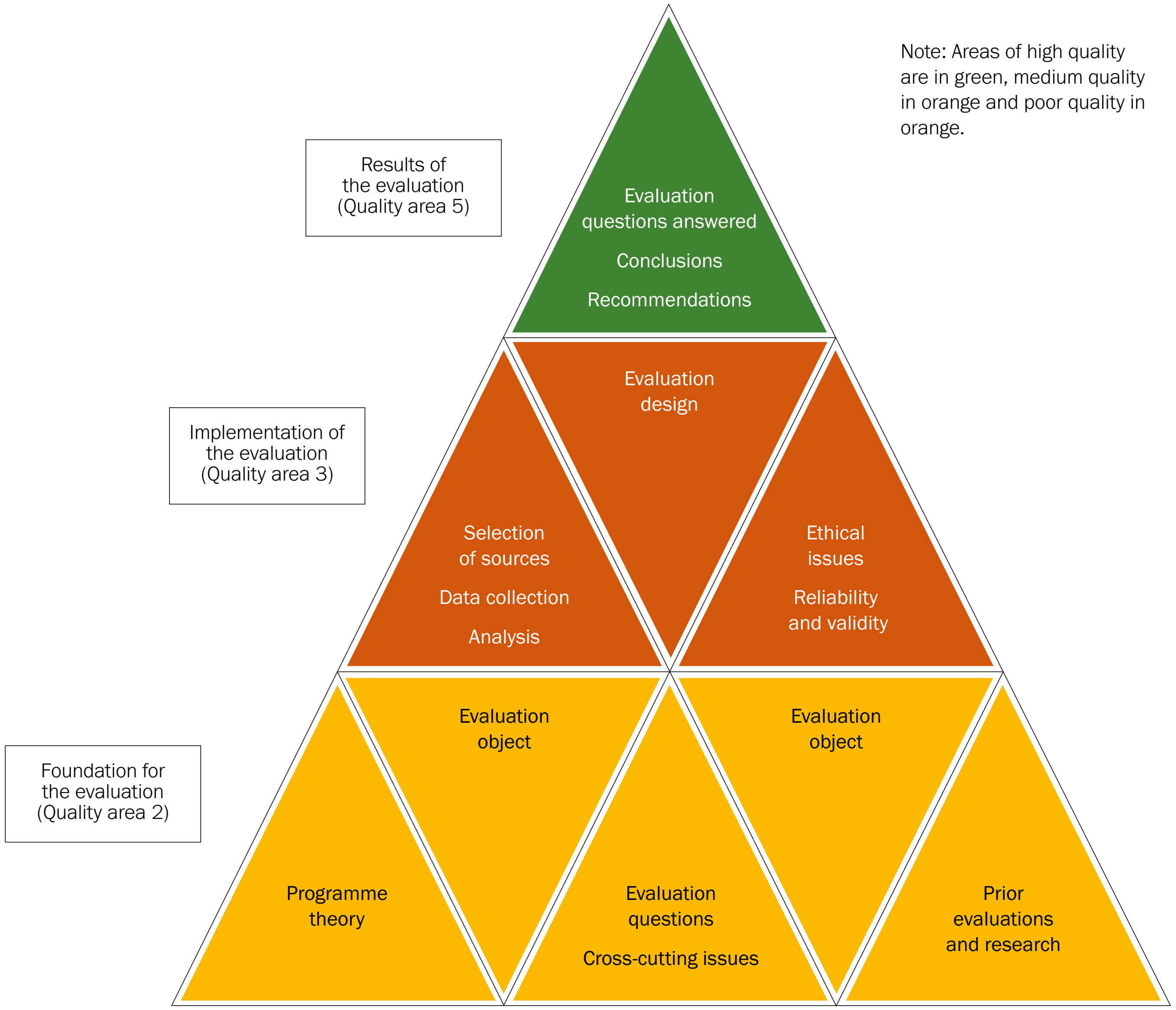
4.1 The foundation for the evaluation

The components along the bottom of the pyramid in Figure 4 set the scene for the evaluation and constitute the foundation for implement the evaluation. Most evaluation reports describe the programme to be evaluated (evaluation object) and evaluation criteria and mention limitations in terms of time and logistics. However, there is less information about context, programme theory and stakeholders and very little on prior reviews or evidence. These factors affect how the evaluation can and should be implemented, and they need to be described so that readers can bear them in mind when interpreting the findings, conclusions and recommendations.

Finding 6: The context is poorly described in both terms of references and reports.

Two-thirds of the terms of references and half of the reports give no or insufficient information about the context of the programme to be evaluated. Most commonly, the terms of references present a brief history of the intervention, which may include some information about the reason the programme was started, and through this, a hint at the context. Few terms of references mention previous evaluations.

Figure 4. Foundation, implementation and results of the evaluation



Note: Areas of high quality are in green, medium quality in orange and poor quality in orange.

This makes it difficult for evaluators to take context and previous reviews into consideration in bidding and in the early stages of planning the evaluation process.

As Example 1 illustrates,[25] the evaluation reports vary widely in how well they describe the context, ranging from reports that do not provide any contextual information at all to those featuring extensive descriptions that take up a large part of the report. Several reports and terms of references come across as targeting an audience that is already familiar with the programme. This has several negative consequences: It limits the reader's ability to assess if the approach, methods and recommendations are appropriate and relevant for the setting; it decreases utility and learning, as similarities and differences to other programmes cannot be assessed; and finally, not all insiders may have the same knowledge and understanding of a programme (e.g. due to staff turnover).

EXAMPLE 1.
TWO APPROACHES TO DESCRIBING CONTEXT

Relevant contextual information is provided, giving the reader a good understanding of the fisheries sector and crimes in South Africa and the region, key stakeholders, main problems etc. *Rater comment, score 4.*

Additional contextual information, setting the scene for the outsider, would have increased the reader's understanding. The report comes across as targeting those already very familiar with the programme. *Rater comment, score 2.*

Finding 7: Few reports and terms of references mention or use information from previous evaluations.

Previous evaluations or research are mentioned in 42 percent of the terms of references and 26 percent of the reports. However, even when referred to, information from these is rarely used to plan the evaluation or to follow up on previous recommendations. This could imply that information that is available is not used, and that data collection may be duplicated. If so, the infrequent use of such information also contributes to evaluation fatigue and waste of resources.[26]

For example, one terms of reference noted that the programme built on two previous grants from the same donor to a similar programme. A closer look at reports on or evaluations of these projects could have provided valuable input to the evaluation, both in terms of data and in terms of paying attention to areas that may have been problematic in those projects. There were no

[25] The raters have provided explanatory comments to all their scores. The rater comments in the textboxes with examples have been slightly edited to ensure consistency and readability. Their content has not been changed.

[26] The team did not attempt to verify if prior evaluations or other research existed, and very few reports or terms of references mentioned that there were no previous evaluations.

EXAMPLE 2.
GOOD USE OF A PREVIOUS EVALUATION

The terms of reference state that the final evaluation is intended to build on a mid-term review that was carried out the previous year. The mid-term review determined the extent to which the outcomes and outputs were met during the previous project period, and the final evaluation should use it as a baseline. The evaluation report has clearly taken this into consideration and makes frequent references to the mid-term review. *Rater comment, score 4.*

signs in the report that this had been done. One terms of reference that is a good example of how previous evaluations can be used is highlighted in Example 2.

Finding 8: Evaluation questions are developed by commissioners. Very few reports discuss, comment on or justify the evaluation questions, and many reports do not at all present the evaluation questions.

All but one terms of reference specify the evaluation questions that the evaluation is expected to respond to. The evaluation questions are mainly clustered around evaluation criteria (effectiveness, efficiency, etc.), but in some cases, they are grouped by project outcomes. In half of the terms of references, the evaluation questions are both clear and relevant and customised to the project. Some of these terms of references have detailed and quite extensive lists of evaluation questions, while others leave the fine-tuning to the evaluators. Very few commissioners leave it to the evaluators to develop the evaluation questions.

Even though evaluation reports should be standalone products, many of the reports assessed by the team do not present the evaluation questions, let alone

EXAMPLE 3.
CONSEQUENCES OF BLINDLY FOLLOWING THE EVALUATION QUESTIONS

One report states, “The structure of this report directly follows the requirements of the ToR and includes the following sections” and goes on to list the main evaluation areas as defined in the terms of reference. The report then uses these as chapter headings, and the questions specified in the terms of reference as sub-headings. The report does not clearly identify or discuss the evaluation questions. This may have revealed that there is some overlap in the terms of reference, and helped give a clearer focus to the evaluation. *Rater comment, score 2.*

discuss, justify or comment on them. When evaluation questions are not presented, it is difficult for the reader to know what the report aims to find out and if it uses relevant methods to do so. The team managed to get access to terms of references for nearly all reports and could verify that in most cases, the reports did respond to the evaluation questions presented in the terms of references. Several reports used evaluation criteria and evaluation questions to structure the report. This made the reports easy to read and the responses to evaluation questions were easy to find. However, if the evaluation questions are not scrutinised by the evaluation team, the result may be the opposite, as illustrated in Example 3 (previous page).

Finding 9: Several reports describe or replicate results frameworks, but few reports make a thorough assessment of the programme theory.

The most common evaluation question refers to whether intended results have been achieved (the evaluation criteria "effectiveness" was included in all 27 evaluations). This is often assessed by comparing planned and actual results, using pre-specified indicators where these exist. However, intended results are often not specified in a way that is easy to measure, and indicators are often missing. In addition, this way of assessing effectiveness focuses on if intended results were achieved, not how they were or were not achieved or the extent to which the programme contributed to results. While quantitative methods, e.g. experimental or quasi-experimental approaches, can be used to prove causality, these require resources and baseline data that are rarely available.

For these reasons, evaluators often rely on qualitative methods to assess effectiveness and causality. By using a theory-based approach, it is possible to discuss causality and assess whether a programme is likely to achieve its intended outcomes and impact.[27] The programme theory[28] is used as point of departure. It describes how the programme is intended to work; it describes how inputs and activities are intended to transform into outputs, outcomes, and impact, and the assumptions and conditions required for this to happen. Despite this, few reports and terms of references describe the programme theory or logic in any detail.

This is sometimes blamed on the lack of a documented programme theory for the programme. However, many terms of references clearly specify that if a documented theory of change or programme theory is not available, the implicit theory of change shall be reproduced by the evaluators.

It is common for the reports to present a results chain consisting of intended impact, intended outcome and outputs but without an explanation of the links and causal pathway between the elements in the chain. In the quality assessment, the raters allowed for a large variety of presentations of programme theories, including results frameworks, log frames, narrative descriptions, diagrams, etc. Despite this, only 40 percent of the evaluation reports provided a satisfactory description of the programme theory. Six reports made a thorough assessment of the programme theory, including analysis of underlying assumptions, but none discussed whether the programme theory was based on previous evaluations or research. Example 4 describes how a programme theory can be reconstructed by the evaluators even when it is missing or of poor quality.

[27] See, e.g., Chen and Rossi (1983). Evaluating With Sense: The Theory-Driven Approach. Evaluation Review, Volume: 7, issue: 3, pages: 283-302.

[28] Also referred to as programme logic, theory of change, intervention logic, etc.

Some terms of references specifically asked for an assessment of the programme theory. In several cases, this was not done: Instead, the reports may describe inputs, outputs, and outcomes, or compare indicators and goals with reported achievements, but without attempting to explain or assess how the project intended to achieve its intended goals or discussing the underlying assumptions. This implies that in several evaluations, even when specifically requested, there is no attempt at explaining how or why the programme succeeded or not. An illustration is presented in Example 5.

4.2 Evaluation design and methods

The components in the foundation of the evaluation affect the methods available to evaluators for collecting evidence, selecting sources and analysing data. Hence, they also affect the quality of evidence, findings, conclusions and recommendations – i.e. the credibility of the results of the evaluation. Without knowing the project, its context and logic as well as the questions to be answered, the reader cannot assess if the selected design and methods are likely to have

EXAMPLE 4. RECONSTRUCTION OF PROGRAMME THEORY

The programme theory is discussed at some length in the report. The design and implementation of the project was found to be inadequate and the evaluators go to some length to reconstruct, revise and improve on the programme theory. *Rater comment, score 4.*

A Theory of Change of the programme, as conceptualised by the review team, is presented and used for assessing the program and the Norwegian support. *Rater comment, score 4.*

EXAMPLE 5. INCOMPLETE UNDERSTANDING OF THE PURPOSE OF ASSESSING THE PROGRAMME THEORY

The programme theory is presented but not assessed, or used to guide the analysis. Although the structure of the theory of change is described, the report does not describe how the programme objectives are intended to be achieved. The report notes that there is a large number of outcomes and lists these in an annex, but does not illustrate the link from the outcomes to the three objectives. The conclusions drawn in the report indicate that the theory of change did not work. *Rater comment, score 2.*

produced robust and evidence-based conclusions and recommendations. If the approach and methods are poorly described, the reader is even more in the dark as to whether conclusions and recommendations can be trusted and what other interventions they may be applicable to.

Evaluation reports therefore need to describe clearly how the evaluation was conducted and why. They need to discuss how credible the presented evidence and derived findings are and identify any remaining limitations that affect the results of the evaluation. Such clarity was missing in many of the quality assessed reports.

Finding 10: Most reports do not explain why evaluators decided to implement the evaluation the way they did. Several reports do not at all describe the evaluation design.

In 59 percent of the reports, the description and justification of the overall evaluation design – i.e. how the chosen package of data and methods intends to contribute to responding to the evaluation questions – is less than satisfactory. Only eleven reports both describe and justify the overall design; six reports do so clearly, giving the reader a good understanding of how the evaluation has arrived at responses to the evaluation questions.

The evaluation design describes how the chosen package of data and methods intends to contribute to responding to the evaluation questions. Some reports present this by listing names of methods, methodologies or approaches. Others present an evaluation matrix listing evaluation issues and questions along with details about data needed to respond to the questions, relevant sources, and methods to collect and analyse the data. Very few of the reports include an evaluation matrix, although some refer to it as part of the inception report. An illustration of how design and methods were presented in two reports is provided in Example 6.

EXAMPLE 6. TWO DESCRIPTIONS OF DESIGN AND METHODS

There is no description of how the evaluation was conducted, neither overall design nor methods are presented. The only information relating to implementation is the affiliation of the evaluators. *Rater comment, score 1.*

The approach and design is very well described, including the motivation for it. The design is illustrated in a figure and tables describe how the different components or phases of the evaluation are linked. There are clear links between evaluation criteria, evaluation questions and methods in an evaluation matrix. The methodology annex of 7 pages gives a very good description for how the review was undertaken, including both methods, distribution of work, analysis and synthesis. *Rater comment, score 4.*

Finding 11: Many reports do not describe how data were analysed or how sources of information were selected. In several cases, the terms of references included lists of whom to interview.

Less than half of the reports provide a satisfactory description of methods used. While most reports mention how data were collected, methods for analysing data are often not described. Even fewer reports describe how the sources of information were selected. Not knowing how the sources of information were selected and how data were analysed makes it impossible for the reader to assess if the evidence is unbiased and if findings and conclusions are credible.

Several terms of references state or suggest the methods that shall be used by the evaluators. Some of these go further and list persons that they want the evaluators to interview. Some authors recognise this as a source of bias, but many do not. Example 7 describes how selection and bias were treated in three reports.

Only four reports (15 percent) provide a clear presentation of how they selected sources, collected data and analysed the information. Very few reports describe in any detail how interviews were carried out or present the tools used for data collection (e.g. interview guides or survey questions). One report's level of detail is exemplary: It describes how the team conducted the interviews and took notes and how the interview notes were used to develop findings. In addition, the interview questions are annexed to the report.

EXAMPLE 7.
TREATMENT OF BIAS AND SELECTION OF SOURCES

The report describes how data was collected and analysed, but sampling or selection methods are not mentioned. As 8 of the 13 interviewees were implementing agency staff, it seems that the risk for biased answers was not considered. *Rater comment.*

The report did not describe how interviewees or documents were selected, although the terms of reference mentions that [the partner organisation] would arrange this prior to the reviewer's arrival. The report did not comment on the risk for bias that comes with letting the evaluated organisation select the sources of information. *Rater comment.*

Norad provided documents and names of interviewees. In the Limitations section, the report describes the potential bias and how this was mitigated through triangulation. *Rater comment.*

Finding 12: Many reports do not provide sufficient references for the sources of the evidence that are presented.

In 59 percent of the reports, the sources of evidence are not consistently provided and many references were missing, making it impossible to link the evidence that is presented to a source. Nearly all reports include lists of interviewees, but in many reports, there are very few references to interviews. This makes it impossible for the reader to know if, or which, interviews were reflected in findings and conclusions. Several reports couched their referencing in general terms, such as "interviews" or "project documents", rather than stating the number of interviews or type of stakeholder or the name of the document. In practise, this means that many reports do not provide enough information for the reader to know what evidence the evaluation has used or if the evidence is credible and unbiased.

Finding 13: The credibility of evidence is poorly described and assessed in the evaluation reports. Ideally, all reports should include a critical assessment of validity and reliability and a clear presentation of limitations that arise from the way the evaluation was implemented. Without this information, it is not possible to assess if conclusions and recommendations are credible and rest on evidence-based findings. However, only three of the reports describe reliability and validity clearly, discussing potential shortcomings in the data collection and analysis and how this affects the findings of the report. Five reports have limited but acceptable discussions of reliability and validity, and six reports make comments that can be interpreted as relating to the issues. In almost half of the reports, there is no critical assessment at all of the quality of data.

Similarly, very few reports contain a discussion of the limitations arising from the way the evaluation was conducted. Half of the reports do not mention limitations at all and in the reports that do describe limitations, the focus is often on limitations to the evaluation process, such as logistical challenges or limited time or budget. Few reports go beyond this and comment on the implications of these limitations for the findings of the evaluation. Only five reports describe limitations that arise as a consequence of the way the evaluation was conducted, such as a risk of biased evidence if not all stakeholder groups could be interviewed, or if interviewees were selected by the commissioner of the evaluated organisation. See Example 8.

EXAMPLE 8.
TREATMENT OF LIMITATIONS

The low response rates, few responses and risk of biased answers in the survey is discussed. The difficulty of assessing contribution is described, and that assessments therefore rely on the evaluator’s professional judgement. The discussion is better than in many reports, but still lacks the link to the credibility of the evidence. *Rater comment, score 3.*

Limitations arising from the evaluation design are clearly described, together with recommendations for how to amend these shortcomings before making decisions about the programme. *Rater comment, score 4.*

4.3 Cross-cutting issues, ethical considerations and quality assurance

Certain aspects should always be considered in evaluations: Norwegian development policy identifies four cross-cutting issues that are to be taken into consideration in all aspects of Norwegian development policy and aid. Similarly, as stated in the Quality Standards for Development Evaluation (OECD, 2010), evaluation ethics should always guide the evaluation process.[29] Despite this, both cross-cutting issues and ethical considerations get very limited attention in the evaluations assessed.

Finding 14: Cross-cutting issues are poorly covered in both terms of references and reports.

As discussed in Chapter 3, very few evaluation commissioners specify that cross-cutting issues shall be assessed. When this is mentioned, it is often in the form of a single evaluation question stating that the evaluation shall assess if the project has taken cross-cutting issues into consideration. Assessing the effect of the evaluated programme on cross-cutting issues, or taking a more extensive approach prescribing, for example, gender-sensitive methods or gender-disaggregated data in the evaluation process, is rare.[30]

The way cross-cutting issues are treated in terms of references and reports gives the impression that these are not considered to be important.

EXAMPLE 9. TREATMENT OF ETHICAL ISSUES

A separate section of the report is devoted to describing the ethical issues and how these were addressed. Ethical issues are clearly stated, e.g. the necessity to maintain confidentially regarding subjects discussed. *Rater comment, score 4.*

Finding 15: Evaluation reports and terms of references do not pay attention to ethical issues.

Ethical issues are mentioned in only five terms of references (20 percent); the remaining 80 percent do not at all mention ethical issues. Of the five, only three discuss it sufficiently to merit a score 3 or 4. This implies that only 12 percent of the terms of references make sufficiently clear requests of evaluators to consider ethical issues in their implementation of the evaluation. However, these requests are often ignored: In four of the five evaluations where the terms of references mentioned ethical issues, the reports did not. In total, only four evaluation reports describe ethical issues and safeguards. Two of these reports scored 4. It is worth noting that these were among the reports with highest overall quality. Example 9 describes one of them.

Whether this treatment of ethical issues reflects neglect in implementation or merely in reporting is unclear. However, the reports contain several examples of ethical issues that were not correctly treated or where a description of ethical safeguards would have been highly relevant. One example is an evaluation of a programme component that targeted violence against women and children. A key part of the methodology was group

[29] OECD (2010), p. 6: “Evaluation abides by relevant professional and ethical guidelines and codes of conduct for individual evaluators. Evaluation is undertaken with integrity and honesty. Commissioners, evaluation managers and evaluators respect human rights and differences in culture, customs, religious beliefs, and practices of all stakeholders. Evaluators are mindful of gender roles, ethnicity, ability, age, sexual orientation, language and other differences when designing and carrying out the evaluation.”

[30] See, e.g., United Nations Evaluation Group (2014). Integrating Human Rights and Gender Equality in Evaluations. New York: UNEG.

discussions with stakeholders where one of the issues discussed referred to experience of abuse and violent behaviour. Despite this, no ethical considerations or potential consequences to the interviewees and others involved are mentioned in the report.

Finding 16: Very few evaluations mention quality assurance.

Only three terms of references make sufficiently clear requests for quality assurance to score 3 or 4; 75 percent of the terms of references do not mention quality assurance at all. None of these three refer to external quality assurance and no details are provided. They merely state that the evaluation shall adhere to the quality standards of the commissioner, implementing organisation or the OECD Development Assistance Committee. No criterion in the rating manual refers to quality assurance of reports, hence this was not rated. However, the team encountered very few reports that describe a quality assurance process. One of these is highlighted in Example 10.

EXAMPLE 10.
QUALITY ASSURANCE OF A REPORT

The terms of reference stated that OECD Development Assistance Committee Quality Standards for Development Evaluations were to be used as a point of reference and specified that these should be mentioned in the inception report. The report adheres to this, noting that “The latest OECD/DAC Quality Standards for Development Evaluations were used as reference, in particular the following quality standards:

- All findings and conclusions are backed by reference to evidence (sources) and the quality and representativeness of the data that the review team bases its findings on is addressed and described;
- Ethical standards (including confidentiality of informants if they request, sensitivity and respect to stakeholders in the [...] project, and the professional conduct of the reviewers) have been adhered to, and interviewees are not generally identified when cited;
- The Do No Harm Principle was observed, and no risk of the [organisation] doing any harm was identified.”

Rater observation regarding quality assurance.

Conclusions

The purpose of this assignment was to contribute to improving the quality of reviews and decentralised evaluations commissioned by the Norwegian aid administration. This study presents clear evidence that there is indeed a need for improved quality. Its discussion of key areas aims to illustrate some of the more serious shortcomings and potential consequences of these findings. This chapter elaborates main conclusions and final remarks.

The study questions contained in the terms of reference for this assignment focus on identifying the main strengths and weaknesses of decentralised evaluations and to what extent these evaluations are based on data, methods and analyses that are likely to produce credible information about the programmes and their outcomes. As noted in Chapter 1, a third study question, related to the main findings of the decentralised evaluations, had to be cancelled due to the low quality of reports.[31]

Conclusion 1: The principal conclusion is that the overall quality of the decentralised evaluations is consistently low.

A large share of the reports had inadequate or poor quality on a majority of the quality criteria, and no report had adequate quality on all criteria. The results are very similar to the two previous quality assessments of decentralised evaluations of Norwegian development cooperation (Norad 2020 and Norad 2017), indicating that there has not been an improvement over time.

Conclusion 2: The main strength of the decentralised evaluations is that they are good at answering evaluation questions and presenting conclusions and recommendations.

The average quality is high on criteria relating to the results of the evaluation. The reports make clear recommendations that respond to the purpose of the evaluations, the recommendations are mainly based on conclusions and conclusions can be linked to findings. Similarly, a substantial share of the evaluations responds to the evaluation questions.

Conclusion 3: Their main weakness is that the reports do a poor job of explaining how they arrived at the response to evaluation questions, conclusions and recommendations.

The average quality is lowest on criteria that relate to implementation of the evaluation. The quality is low on presentation of evaluation design, sources of evidence, methods for selecting data and analysis of data. When it comes to linking findings to evidence and evidence to sources, the quality is less than satisfactory in most of the reports.

Conclusion 4: Many reports lack sufficient information about the evaluated programme and context.

There are major shortcomings in descriptions of the foundations for the evaluation, i.e. context, the evaluated programme and how the programme is intended to achieve its goals (the programme theory).

[31] The third study question has been replaced by summaries of three evaluation reports of high quality. These are presented in Annex 6.

Conclusion 5: Many reports lack sufficient information to allow the reader to assess if findings, conclusions and recommendations are based on credible evidence.

Several findings in this quality review indicate that readers will find it difficult to trace findings back to evidence and sources and to assess if the decentralised evaluations present credible information about the programmes and their outcomes. Several reports lack references to sources or have a biased selection of sources.

Conclusion 6: There is a lack of transparency about limitations and quality of evidence.

The quality was particularly low on quality criteria related to reliability and validity and to limitations resulting from the way the evaluation was conducted. Few reports discuss these issues and when they do so, the discussion is often superficial.

5.1 Final remarks

Poor quality can result in harmful decisions and missed opportunities for learning

The combined effect of the aforementioned strengths and weaknesses is that while most reports express clear conclusions and recommendations, many evaluations leave the reader in the dark as to how these were arrived at. This implies that many recommendations about funding and programming are not based on evidence or are based on evidence that may not be credible. This may have several effects. The present donor, project manager or desk officer cannot know if they can trust the report's conclusions and recommendations, for instance, and if recommendations are not based on credible evidence they may result in decisions that harm the project or beneficiaries. Future project staff and managers who do not have full information about the programme will have difficulty interpreting the report findings; hence, the evaluation may not be fully and correctly used and opportunities to draw lessons and improve future programming may be missed.

Without sufficient information about the programme, programme theory and context, it is difficult to determine whether the findings may be applicable to other interventions and if so, which ones. This limits the extent to which the decentralised evaluations can be used for learning in other interventions and contexts. Learning is also limited by the fact that most evaluations are not made publicly available and by the lack of information about evaluations that have been implemented.

Introducing quality assurance measures could improve the quality of evaluations

Very few terms of references include requests for quality assurance, and even fewer reports mention quality assurance procedures. Yet, many of the shortcomings observed by the raters in this quality assessment would have been just as easily captured by a basic quality assurance system. The quality of decentralised evaluations could have been higher if terms of references clearly outlined the expected quality standards; draft reports were required to demonstrate clear linkages between sources, evidence, findings, conclusions and recommendations; and quality assurance requirements were summarised and explained in a basic easy-to-use and easy-to-access quality assurance tool constructed for internal use.

References

Cooney, Rojas, Arsenault and Babcock (2015). Meta-Evaluation of Project and Programme Evaluations in 2012-2014. Evaluation on Finland's Development Policy and Co-Operation, 2015/3.

Department of Foreign Affairs and Trade, Australian Government (2018). Review of 2017 Program Evaluations Prepared by the Office of Development Effectiveness (ODE).

Norad (2020). Quality Assessment of Decentralised Evaluations in Norwegian Development Cooperation (2018–2019). Evaluation Department Report 6/2020, Norad.

Norad (2017). The Quality of Reviews and Decentralised Evaluations in Norwegian Development Cooperation. Evaluation Department Report 1/2017, Norad.

OECD (2019a). OECD Development Co-operation Peer Reviews: Norway 2019, OECD Development Co-operation Peer Reviews, OECD Publishing, Paris.

OECD (2019b), Development Assistance Committee: Better Criteria for Better Evaluation Revised Evaluation Criteria Definitions and Principles for Use, OECD Development Assistance Committee Network on Development Evaluation. OECD Publishing, Paris.

OECD (2016). Development Assistance Committee: Evaluation Systems in Development Cooperation: 2016 Review. OECD Publishing, Paris.

OECD (2010). Quality Standards for Development Evaluation, DAC Guidelines and Reference Series. OECD Publishing, Paris.

OECD (2002). Glossary of key terms in evaluation and results based management. OECD Publishing, Paris.

United Nations Evaluation Group (2014). Integrating Human Rights and Gender Equality in Evaluations. New York: UNEG.

Annex 1. Terms of references

Annual Assessment of the Quality of Reviews in Norwegian Development Cooperation 2019–2021

Multi-year assignment to make running quality assessments of reviews and decentralised evaluations published annually in the years 2019–21, and to summarise their strengths and weaknesses in an annual publication, also presenting the most important knowledge generated from the reviews and decentralised evaluations.

1. BACKGROUND

Reviews and decentralized evaluations[1] of development projects and programmes are an important source of information about the results of Norwegian development cooperation.[2] Credibility and utility of these reviews and decentralised evaluations is therefore important.

Achieving adequate quality of decentralised evaluations is a challenge in many agencies.[3] Therefore, many agencies, both bilateral donors and multilateral organisations, have institutionalised an external quality assessment mechanism to improve quality.[4] Arrangements vary, but most aim to improve evaluation quality both directly by rating quality of commissioned reviews and decentralised evaluations and indirectly by raising awareness about the importance of evaluation quality.

The Norwegian aid administration (MFA, Norad, Embassies) has no quality assessment mechanism for reviews and decentralised evaluations. The assignment will be a first step to establish this.

Reviews in the Norwegian Aid Administration

The quality of reviews and decentralised evaluations commissioned by the Norwegian aid administration[5] has been questioned in evaluations and studies in recent years, the most recent being the study of 2014 reviews and decentralised evaluations, Evaluation department report 1/2017.[6]

[1] Hereafter mainly referred to as ‘reviews and decentralised evaluations’.

[2] The Evaluation Department in Norad is responsible for conducting strategic level evaluations, while these project and programme reviews and decentralised evaluations are the responsibility of the grant manager.

[3] OECD DAC (2016) Evaluation Systems in Development Cooperation: 2016 Review. OECD Publishing, Paris.

[4] E.g. DFAT (2018) ‘Review of 2017 Program Evaluations’, Office of Development Effectiveness, Department of Foreign Affairs and Trade, Australian Government; Independent Evaluation Office (2017) Review of the Quality Assessment of the 2016 Decentralised Evaluations, United Nations Development Programme.

[5] For this purpose, this includes The Ministry of Foreign Affairs, Royal Norwegian Embassies managing ODA-funds and Norad. Norfund and Norec, formally part of the Norwegian aid administration, are not part of this review.

[6] Evaluation department Norad report 1/2017; Evaluation Department Norad Report 1/2014; OECD-DAC peer review 2013; Evaluation Department Norad Report 7/2012; Evaluation Department Norad Report 4/2018.

The Evaluation department report 1/2017[7] found that more than half of the reviews and decentralised evaluations were of inadequate quality in terms of their methodological basis, assessment of results and that findings and conclusions were not sufficiently well founded. The evaluation found that ethical considerations were not adequately covered in the reviews and decentralised evaluations. The evaluation indicates that reviews and decentralised evaluations are highly used by the responsible unit but that the knowledge generated by the reviews and decentralised evaluations and decentralized evaluations is not made available to others.[8]

Guidance for why, when, and how to undertake reviews and decentralised evaluations is given in the GMM and requirements are specified in the rules[9] for each grant scheme. A review, as defined in the Grant Management Manual (GMM)[10] is 'a thorough assessment with focus on the implementation and follow-up of plans', which may be undertaken underway (mid-term review) or after finalisation to assess the effect of the programme/ project (end review).

Reviews are commissioned by the unit responsible for grant management (Embassies, MFA, Norad[11]), implementing partners/grant recipients, and other agencies/co-sponsors. An estimated 60–70 reviews and decentralised evaluations are undertaken per year.[12] All reviews and decentralised evaluations and evaluation reports shall be submitted to the evaluation portal[13], as per grant scheme rules. However, this is currently not common practice, so the number of reviews and decentralised evaluations published in the evaluation portal is likely to be much lower than that, and for 2019 may be as few as 20–30 reviews and decentralised evaluations.

The requirement to conduct evaluations follows from the Regulations for Financial Management in the Government Administration.[14] Accompanying guidance material emphasise systematic use of evaluations as a source of management information and learning.[15]

2. PURPOSE AND OBJECTIVES

The purpose of this assignment is to contribute to improve the quality of reviews and decentralised evaluations and decentralized evaluations

[7] Evaluation department Norad report 1/2017 'The Quality of Reviews and Decentralised Evaluations in Norwegian Development Cooperation'.

[8] This was found in a mapping conducted in preparation for evaluation report 1/2017, Evaluation Department Norad (2015) Study of Reviews and Decentralised Evaluations in Norwegian Development Cooperation – mapping. Report 11/2015.

[9] Grant scheme rules define the objectives, target group and criteria for each grant scheme, as well as requirements for follow up of agreements. Each grant scheme has a separate set of rules, though there are commonalities.

[10] The manual applies to all grants managed by the Ministry of Foreign Affairs (including the Embassies managing ODA-funds) and Norad. Ministry of Foreign Affairs, 'Grant Management Manual. Management of Grants by the Ministry of Foreign Affairs and Norad'. 05/2013. (Not available online.)

[11] Norad, in line with its mandate as quality assurer of Norwegian assistance, will also commission reviews and decentralised evaluations on behalf of Embassies and the Norwegian Ministry of Foreign Affairs, as part of its technical support.

[12] Based on findings of mapping in Evaluation Department Norad Report 11/2015. The number of reviews and decentralised evaluations registered in the Evaluation portal is likely to be much lower. It is expected that this assignment may raise awareness and increase the number.

[13] https://evalueringsportalen.no/

[14] 'Reglement for økonomistyring i staten' (2003) and 'Bestemmelser om økonomistyring i staten' https://www.regjeringen.no/globalassets/upload/fin/vedlegg/okstyring/reglement_for_okonomistyring_i_staten.pdf Ministry of Finance has issued a guide for undertaking evaluations 'Veileder til gjennomføring av evalueringer' (2005).

[15] Strategisk og systematisk bruk av evaluering i styringen. Veileder. Direktoratet for Økonomistyring (DFØ) (2011).

commissioned by the Norwegian aid administration, by giving an annual diagnostic of the quality of reviews and decentralised evaluations published. Furthermore, the purpose is to make knowledge generated in these reviews and decentralised evaluations more accessible by presenting key findings in an annual publication.

The assignment contains both accountability and learning aspects. Main intended users are the Section for Grant Management in the Ministry of Foreign Affairs and Department for quality Assurance in Norad. The quality review will provide these quality assurance units with information about the strengths and weaknesses of reviews and decentralised evaluations commissioned by the aid administration annually, which may be used to take measures to improve quality.

Users also include MFA Departments and Norwegian Embassies managing ODA-funds, Departments in Norad and other parts of the aid administration, as well as partners in Norwegian Development Cooperation. The publication of an annual report may contribute to increase commissioners' and evaluators' attention to quality.

The objectives of the study are to:

1. Assess the quality of reviews and decentralised evaluations of Norwegian development cooperation;

2. Identify strengths and weaknesses of reviews and decentralised evaluations:

3. Summarise findings from the reviews and decentralised evaluations, taking into consideration the credibility assessment made under objective 1.

3. SCOPE OF WORK

The assignment will cover reviews and decentralised evaluations published[16] in the year 2019, with the option of extending to 2020 and 2021-reviews and decentralised evaluations respectively.

The study includes reviews and decentralised evaluations and decentralized evaluations commissioned by MFA, Norad[17] and Norwegian Embassies, that are published in the Evaluation portal of the Norwegian government. The Evaluation portal website is available in Norwegian only.

The consultant will search the portal at least semi-annually to identify relevant reviews and decentralised evaluations and will assess the quality of each single review obtained and accompanying TOR (if annexed). In addition, the consultant may also have to reach out to the sections and departments in the MFA, Norad and Embassies to identify additional reviews and decentralised evaluations. The list of reviews and decentralised evaluations to be assessed and rated must be approved by the Evaluation Department prior to assessment/rating.

The consultant will produce an annual quality assessment report with summary and analysis of

[16] Published on the Evaluation Portal within 31st December each year.

[17] Primarily project, programme and portfolio reviews and decentralised evaluations (mid-term or end reviews and decentralised evaluations or evaluations). If other types of reports are to be included in the scope, this requires prior approval from the Evaluation Department. Thematic, centralized evaluations carried out by the Evaluation Department in Norad are not part of the scope of this study.

the quality of reviews and decentralised evaluations assessed throughout the year, to present conclusions on the overall quality.

The annual quality assessment report will present the most important findings from reviews and decentralised evaluations, taking into consideration the credibility of the findings, as per the quality assessment made. To the extent that the material allows, analysis of review findings across reviews and decentralised evaluations may seek to identify general trends and patterns.

The consultant will assess quality using the quality assessment template (appendix 1) based on the OECD-DAC quality standards[18], developed for the previous evaluation of the quality of reviews and decentralised evaluations commissioned by the Evaluation department (Evaluation department Report 1/2017). Reviews will be rated 1-4 on each quality criterium in the template and a justification will be given for each score. Individual reviews and decentralised evaluations will not be given an average overall rating.

The consultant will calculate average scores for each key quality area for each review (1. Summary, style and structure; 2. Review purpose, objectives, and scope; 3. Methodology; 4. Application of the OECD DAC evaluation criteria; 5. Analysis, data, findings, conclusions, and recommendations), and will provide a comment to substantiate the score.[19] Average scores per key quality area will be used to identify strengths and weaknesses across the whole sample of reviews and decentralised evaluations.

Findings will be compared and discussed against findings from the previous year (2019-reviews and decentralised evaluations may be compared with findings from the assessment of 2014-reviews and decentralised evaluations (in Report 1/2017)).

Quality in this assignment will be understood as quality of the written review report, as measured against the quality assessment template. Emphasis will be on soundness of methodology and analysis, given the weaknesses identified in that regard in previous evaluations. Other aspects of quality such as the quality of review process, use of review findings, and usefulness of the knowledge generated will not be considered. This is a limitation of the study.

The annual assessment report will present descriptive statistics of basic characteristics of the reviews and decentralised evaluations: sector; country or region; commissioning unit (MFA, Embassy, Norad); whether the review is carried out by external consultants, internal team or a mixed team.

4. STUDY QUESTIONS

The following questions will guide the assignment:

1. To what extent are reviews and decentralised evaluations based on data, methods and analyses that are likely to produce credible information about the programmes and their outcomes?

2. What are the main strengths and weaknesses of reviews and decentralised evaluations of Norwegian development cooperation? Assessed per quality area of the template for example.

[18] OECD Development Assistance Committee http://www.oecd.org/dac/evaluation/qualitystandardsfordevelopmentevaluation.htm

[19] As the individual quality criteria will not be weighted, a qualitative comment will allow for a correction where the average numerical score may give a skewed picture. It will also allow for more explanation as needed, since some quality areas encompass a range of aspects.

3. What are the main findings of the reviews and decentralised evaluations in the sample?

5. METHODOLOGY

The study will primarily be carried out as a desk review. Data sources include:

- The evaluation portal (DFØ) (evalueringsportalen.no). In addition, sections, departments and embassies may have to be contacted to retrieve additional reviews and decentralised evaluations.

- Key governing documents such as the MFA Grant Management Manual, rules and guides issued by the Ministry of Finance and the Directorate for Finance Management (DFØ) and other relevant documents.

The assessment of reviews and decentralised evaluations will be made according to the templates in appendixes 1 (Guidance Manual: Quality Assessment Manual for Decentralised Evaluations and Reviews and 2 (Template for Quality Assessment of Terms of References).

The consultant shall outline a strategy to ensure the objectivity, reliability, and validity of review ratings. This could include how to ensure reliability across different raters (inter-rater reliability), or across different reviews and decentralised evaluations for the same rater (inter-report reliability). Limitations to the chosen approach should be described, including strategies to counteract these.

The inception note will include a brief outline of the consultant's understanding of the criteria, including any limitations that the consultant may foresee.

The inception note will also include the consultant's approach to synthesis of the main findings in the reviews and decentralised evaluations in the sample, mindful of the quality assessment, particularly related to methodological weaknesses identified in the reviews and decentralised evaluations.

The annual report shall discuss any limitation to the chosen approach, and include an assessment of the objectivity, reliability and validity of findings.

The consultant may in the annual assessment report for the 2019-review propose adjustments to the assessment tools based on the experience from the first annual volume.

Rating and key characteristics for all reviews and decentralised evaluations in the sample shall be systematized in an Excel database, which shall be the basis for simple statistical analysis and be submitted as a separate deliverable.

The consultant shall discuss relevant ethical issues to the assignment and suggest safeguards to counteract these if needed.

The assignment shall be carried out in accordance with relevant guidelines from the Evaluation Department (available at norad.no/evaluationguidelines).

6. ORGANISATION OF THE ASSIGNMENT

The study will be managed by the Evaluation Department. The consultant will report to the Evaluation Department through the team leader. The team leader shall be in charge of all deliveries and will report to the Evaluation Department on the progress of the assignment, including any problems that may jeopardise the assignment, as early as possible.

All decisions concerning the interpretation of these Terms of Reference, and all deliverables are subject to approval by the Evaluation department.

Quality assurance shall be provided by the institution delivering the services prior to submission of all deliverables.

7. BUDGET, TIME FRAME AND DELIVERABLES

The consultant will be remunerated at two working days per rated review report, and thirty-eight working days for each Annual Quality Assessment Report – for inception work including search for reviews and decentralised evaluations, synthesis, analysis, reporting, presentation and quality assurance.

It includes the following deliverables:

- Annual inception report (not exceeding 5 pages) to be submitted together with a preliminary list of reviews and decentralised evaluations retrieved from the Evaluation Portal;
- Draft Annual Quality Assessment Report (not exceeding 15 pages, excluding summary and annexes) for preliminary approval by EVAL and circulation to the stakeholders. After circulation to the stakeholders, the Evaluation department will provide feedback;
- Database documenting quality scores for all reviews and decentralised evaluations, including written justification, in Excel-format (to be submitted together with the Draft Annual Quality Assessment Report);
- Final Annual Quality Assessment Report, not exceeding 15 pages, excluding summary and annexes;
- Annual seminar/workshop in Oslo to present the Annual Quality Assessment Report.
- Semi-annual list of reviews and decentralised evaluations to be rated, retrieved from the Evaluation Portal (applicable as of the 2020-reviews and decentralised evaluations), to be approved by Evaluation Department

All data, presentations, reports are to be submitted in electronic form in accordance with the deadlines set in the tender document and the Evaluation department's guidelines (available at norad.no/evaluationguidelines). EVAL retains the sole rights with respect to all distribution, dissemination and publication of the deliverables.

Annex 2. Descriptive statistics

Table 1. Average and standard deviation, report quality criteria

Report quality criteria	Average	Standard deviation
1.1 Executive summary	2.48	0.79
1.2 Style and structure	2.89	0.79
2.1 Purpose of the evaluation	2.96	0.96
2.2 Evaluation object	2.93	0.77
2.3 Description of the programme theory	2.52	1.03
2.4 Context	2.70	0.85
2.5 Scope	2.85	0.93
2.6 Evaluation questions	2.07	0.94
2.7 Existing evidence base	2.04	0.84
3.1 Description and justification of the evaluation design	2.44	1.03
3.2 Description of methods	2.56	0.83
3.3 Methodological application	2.41	1.03
3.4 Reliability and validity of evidence	1.93	1.05
3.5 Sources of evidence	2.48	0.63
3.6 Limitations	1.96	1.17
3.7 Ethical issues	1.37	0.91
5.1 Response to evaluation questions	3.04	0.82
5.2 Findings	2.67	0.77
5.3 Conclusions	3.00	0.86
5.4 Recommendations are based on conclusions	2.96	0.82
5.5 Recommendations respond to the purpose of the evaluation	3.28	0.87
5.6 Recommendations are clear and actionable	2.72	0.72
6.1 Overall quality of the report	2.74	0.75

Table 2. Average and standard deviation, terms of reference quality criteria

Terms of reference quality criteria	Average	Standard deviation
1.1 Rationale and purpose of the evaluation	3.50	0.76
1.2 Specific objectives of the evaluation	3.75	0.66
1.3 Context of the development intervention being evaluated	2.08	1.04
1.4 Previous evaluation	1.79	1.12
1.5 Evaluation object	3.25	0.78
1.6 Scope	3.29	0.79
1.7 Evaluation criteria	3.00	0.87
1.8 Evaluation questions	3.42	0.70
1.9 Feasibility	3.13	0.70
2.1 Review process	2.92	0.86
2.2 Deliverables	3.25	0.88
2.3 Quality assurance	1.42	0.81
3.X Human rights	2.00	1.08
3.1 Gender	2.04	1.02
3.2 Climate and Environment	2.00	1.12
3.3 Anti-corruption	1.63	0.95
3.4 Ethics	1.46	1.00
3.5 Expected limitations to the review	1.42	0.64
4.1 Overall rating of the terms of reference	2.67	0.55

List of annexes

Annex 3: Methodology
- Annex 3.1 Evaluation matrix
- Annex 3.2 Ethical considerations, risks and limitations
- Annex 3.3 Request for evaluation reports
- Annex 3.4 Data collection tools

Annex 4: Decentralised evaluations included in the assessment

Annex 5: Data and descriptive statistics

Annex 6: Good practise evaluations
- Annex 6.1 End review of the Tanzanian Agricultural Partnership programme phase II
- Annex 6.2 Mid-term review of the FishFORCE project
- Annex 6.3 End-term evaluation of the Women Pioneers in the Judiciary project

Annex 7: Profiles of the assessment team

Annexes 3 to 7 can be found in a second part of the report on www.norad.no/evaluation

Acronyms and abbreviations

DAC	Development Assistance Committee (OECD)
MFA	Ministry of Foreign Affairs (Norway)
Norad	Norwegian Agency for Development Cooperation
OECD	Organisation for Economic Co-operation and Development
TOR	Terms of Reference
UN	United Nations
UNEG	United Nations Evaluation Group

All reports are available at our website www.norad.no/evaluation

Related Reports

You may also be interested in these related reports from the Evaluation Department

6.2020 Quality Assessment of Decentralised Evaluations in Norwegian Development Cooperation (2018–2019)

2.2020 Evaluation of the Norwegian Aid Administration's Practice of Results-Based Management

4.2018 Evaluation of the Norwegian Aid Administration's Practice of Results-Based Management/

1.2017 The Quality of Reviews and Decentralised Evaluations in Norwegian Development Cooperation

4. 2015 Experiences with Results-Based Payments in Norwegian Development Aid

5.2015 Basis for Decisions to use Results-Based Payments in Norwegian Development Aid

1.2014 Can We Demonstrate the Difference that Norwegian Aid Makes?

8.2012 Use of Evaluations in the Norwegian Development Cooperation System

Funded by the Evaluation Department:
(OECD-DAC) Evaluation systems in development co-operation OECD-DAC 2016 Review

Norad

EVALUATION DEPARTMENT

Norwegian Agency for
Development Cooperation

www.norad.no
post-eval@norad.no

Cover photo Ken Opprann
ISBN 978-82-8369-071-2
October 2021