

EVALUATION DEPARTMENT

Report 6 / 2020

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



Commissioned by
The Evaluation Department

Carried out by
Ternstrom Consulting AB

Written by
Ingela Ternström (team leader and main author),
Jock Baker, Stefan Dahlgren, Eva Lithman,
Abid Rehman and
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 2020



Content

Foreword	4	4. Discussion and Conclusions	46
Acknowledgements	5	References	49
Executive Summary	6	Annex 1: Terms of References	50
1. Background and Purpose	8	Annex 2: Data Collection Tools	56
2. Approach and Methodology	11	Annex 3: Presentation of Data	66
2.1 Quality Assessment Tool	11	Annex 4: Best Practise Evaluations	75
2.2 Identification and Selection of Decentralised Evaluations	13	List of Annexes	87
2.3 Scoring Process and Data Analysis	14	List of Tables and Figures	88
2.4 Limitations	15	Acronyms and Abbreviations	89
2.5 Ethical Considerations	16		
3. Findings: The Quality of Evaluation Reports and Terms of References	17		
3.1 Findings and Distribution of Scores by Report	17		
3.2 Report Quality Criteria Averages	20		
3.3 Report Quality Criteria Scores	22		
3.4 Terms of Reference Quality Criteria	42		



Foreword

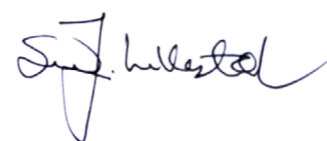
Most evaluations commissioned by the Norwegian Aid Administration are initiated by the units responsible for grant management in the development aid administration. These evaluations - commonly called decentralised evaluation or reviews - are intended to form a key part of the evidence base for documenting 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. This study reveals that the overall quality of decentralised evaluations is still low.

For decentralised evaluations to fulfil its intension, 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 evaluation was carried out by the Swedish consultancy company Ternstrom Consulting AB and we thank the team for a job well done.

Oslo, October 2020



Siv J. Lillestøl

Acting Director, Evaluation Department



Acknowledgements

This report summarises the results of a quality assessment of a large number of reviews and decentralised evaluations. The team is grateful to have the opportunity to share the information provided in all these reports. Although the main focus has been on descriptions, methodology and other technical aspects, the team has been impressed by the richness of information presented in the reports and by the work done by evaluators and in the evaluated interventions.

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. The report was prepared by the team leader with assistance from other team members.



Executive Summary

Decentralised evaluations of development projects and programmes are an important source of information about the results of Norwegian development cooperation. Credibility and utility of these decentralised evaluations is therefore important. An evaluation published by the Evaluation Department in 2017 found that the quality of these decentralised evaluations was poor, questioning the credibility of findings and conclusions. This study is a follow-up of the 2017 evaluation. The study assesses the quality of decentralised evaluations, commissioned by the Ministry of Foreign Affairs, Norad and embassies, published in 2018-2019. A second follow-up study will be done in 2021 looking at evaluations published in 2020.

The purpose of the follow-up studies 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, the study provides information about the credibility of information presented in decentralised evaluations. Finally, the study may also increase commissioners' and evaluators' attention to quality in general.

Originally the study was set up to also summarise findings arising from these evaluations. However, due to the low quality of the reports in this study, methodology and findings from three high-quality reports is presented instead.

A standardised rating manual was used to assess quality. The quality criteria set out in the rating manual are largely based on the OECD DAC quality standards and quality criteria for evaluation, as well as cross-cutting themes defined by the Norwegian Ministry of Foreign Affairs. Extensive measures were taken to ensure consistent assessment, including double scoring of nearly half of the evaluation reports. The quality assessment is limited to information presented in written reports and terms of references. Other aspects of the evaluation process are not included in the study.

The study reveals that the overall quality of decentralised evaluations is low. Nearly half of the reports and the terms of references had poor or less



than adequate quality on more than half of the quality criteria.

Several aspects of evaluation quality were judged to be poor. Approach and methods for collecting and analysing data were often very briefly described. Few reports explained why specific methods and sources were selected, and critical reflection and transparency regarding the quality of data was rare. Other weaknesses are that assessments of efficiency were of poor quality, ethical and anti-corruption issues were mainly ignored and mandates (terms of references) were not always adhered to.

While style and structure of reports, as well as description of the assignment, tend to be stronger, this cannot make up for existing weaknesses above: The assessment of evaluation reports indicate that a large number of decentralised evaluations are not based on data, methods and analyses that are likely to produce credible information about the programmes and their outcomes.



Background and Purpose

This report presents the findings of an independent quality assessment of reviews and decentralised evaluations commissioned by the Norwegian aid administration¹ and published during 2018–19. The Evaluation Department in Norad commissioned this assignment as part of an effort to improve the quality of reviews and decentralised evaluations.

The 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, and quality assurance was provided by Mr Abhijit Bhattacharjee. Together, the team members have commissioned, carried out, quality assured, quality assessed or managed a substantial number of evaluations and reviews. 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.²

During the two years covered by this report, Norwegian aid amounted to a total of 72 billion Norwegian kroner.³ Evaluations are an important source of information about both results and the implementation methods of the many organisations and interventions that apply for funding, and many evaluation reports make explicit recommendations regarding funding. Programme managers also use the information in decentralised evaluations, which provide an opportunity to capture problems, make revisions and identify good practises. Decentralised evaluations thus potentially can affect both funding and implementation in a large number of interventions, with extensive impact on implementing organisations and target populations worldwide.

To ensure credibility and utility, it is crucial that the data, evidence and findings that form the basis for conclusions and recommendations are of high quality and give a correct, reliable and unbiased picture of reality. This requires the methodology used to collect and analyse information to be of high quality, and that shortcomings be presented in a transparent manner.

At the same time, achieving adequate quality of decentralised evaluations is challenging for the Norwegian aid administration, just as it is for many other agencies.⁴ This was illustrated in the 2017 report, *The Quality of Reviews and Decentralised Evaluations in Norwegian Development Cooperation*.⁵ This evaluation found the decentralised evaluations undertaken in 2014 to be of inadequate quality,

1 The aid administration here refers to the Ministry of Foreign Affairs, Norwegian embassies and Norwegian Agency for Development Cooperation (Norad).

2 Annex 7 presents short descriptions of the consultants.

3 See <https://norad.no/om-bistand/norsk-bistand-i-tall/>. Norwegian development aid was 34 635 million Norwegian kroner in 2018 and 37 764 in 2019.

4 OECD DAC (2016). *Evaluation Systems in Development Co-operation: 2016 Review*. OECD Publishing, Paris.

5 Evaluation Department (2017). *The Quality of Reviews and Decentralised Evaluations in Norwegian Development Cooperation*. Evaluation Department report 1/2017. Oslo: Norad. Authored by Itad Ltd. and Chr. Michelsen Institute.



particularly in terms of methodology and assessment of results. Findings and conclusions were not sufficiently well-founded, and ethical considerations were not adequately covered. The authors also noted that the responsible units nevertheless used these decentralised evaluations.

The present assignment aims to contribute to improving 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. Another aim is to make the knowledge generated by these reviews and decentralised evaluations more accessible by presenting key findings in an annual publication.⁶ A final goal is to contribute to both accountability and learning. Three objectives are identified in the Terms of Reference:

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

2. Identify strengths and weaknesses of reviews and decentralised evaluations.
3. Summarise findings from the reviews and decentralised evaluations, taking into consideration their credibility and based on the assessed quality.

Due to the low quality of the assessed reviews and decentralised evaluations, the third objective was revised and instead, best practice in terms of quality is presented for three reports that received the highest scores.⁷

The assignment incorporates accountability and learning aspects. The main intended users of the report are the quality assurance units of the Ministry of Foreign Affairs and Norad, which may use the information about strengths and weaknesses of reviews and decentralised evaluations to adopt measures to improve quality. Other parts of the Norwegian aid administration may also use the information, and it is hoped that the information presented in this and subsequent annual reports may contribute to increase commissioners' and evaluators' attention to quality.

While reviews and decentralised evaluations commissioned by the Norwegian aid administration are the object of the assignment, its scope is decentralised evaluations that meet the following criteria:⁸

- finalised during 2018–19
- midterm or end reviews or decentralised evaluations of projects or programmes funded via Norwegian development cooperation
- commissioned by Norad, the Ministry of Foreign Affairs or Norwegian embassies
- carried out by internal or external teams

It should be noted that the scope only includes terms of references and evaluation reports. Tenders, inception reports and other aspects of the evaluation process are not included. The quality assessments shall be made using the same tools as in Evaluation Department (2017). A document review of other assessments of

⁶ See the Terms of Reference for the assignment in Annex 1.

⁷ This change was done in agreement with the Evaluation Department in Norad.

⁸ These criteria were agreed upon during the inception phase. The Terms of Reference specified 2019 as the time period, but this was broadened to 2018–19 due to the small number of reviews and decentralised evaluations available online.



evaluation quality, including for example OECD⁹ peer reviews and other donors' assignments, provided supplementary information regarding approach, methodology and tools.¹⁰

The requirement to conduct evaluations follows from the Regulations for Financial Management in the Government Administration (Økonomiregelverket). The same regulations apply to both reviews and decentralised evaluations, and the term “decentralised evaluations” is used henceforth for both reviews and decentralised evaluations. Occasionally, the terms *evaluation* and *evaluation report* are used for ease of reading.

Chapter 2 of this report describes the approach and methodology used by the assessment team in the quality assessment; complementary information is presented in Annexes 2 and 5. Chapter 3 summarises findings from the quality assessment of evaluation reports and terms of references; Annex 3 presents data and Annex 4 presents summaries of the three reports with highest scores. Chapter 4 summarises the team's findings and conclusions drawn. Part II of the report contains annexes with additional information about methodology (Annex 5), the quality assessed evaluations (Annex 6) and the assessment team (Annex 7). Part II of the report can be found on <https://www.norad.no/evaluation>.

9 Organisation for Economic Co-operation and Development.

10 See, for example, OECD DAC (2016). *Evaluation Systems in Development Cooperation: 2016 Review*. OECD Publishing, Paris; OECD (2019). *OECD Development Co-operation Peer Reviews: Norway 2019*, OECD Publishing, Paris, <https://doi.org/10.1787/75084277-en>; Department of Foreign Affairs and Trade, Australian Government (2018). *Review of 2017 Program Evaluations Prepared by the Office of Development Effectiveness (ODE)*; 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.



Approach and Methodology

This assignment is a standardised assessment of the quality of reports and terms of references of decentralised evaluations¹¹ using a strict, predefined tool. The team's main focus was developing an approach that ensures the tool is consistently applied and makes the quality assessment process as accurate as possible. The approach, summarised in a brief evaluation matrix in Annex 5, includes:

- a system and process of scoring that contribute to consistent use of the scoring tool across raters¹² and evaluation reports and over time, i.e. to reduce bias

- a logical and well-functioning structure for retrieving and storing reports, collating data (report data, quality data and findings), and analysing data.

2.1 Quality Assessment Tool

The tool used to assess the quality of decentralised evaluations and terms of references was prepared by Itad Ltd. and Chr. Michelsen Institute in Evaluation Department (2017), and is based on the OECD DAC¹³ quality standards for evaluating development assistance.¹⁴ It consists of the following:

- A Guidance Manual with scoring instructions for 34 quality criteria divided into five quality areas:¹⁵
 - summary, style and structure
 - evaluation purpose, objectives, and scope
 - methodology
 - application of OECD DAC evaluation criteria
 - analysis, data, findings, conclusions, lessons learned, recommendations and cross-cutting issues.
- A quality assessment template for 18 quality criteria for terms of references, divided into three quality areas:
 - evaluation purpose, objectives, object and scope
 - evaluation process and quality assurance
 - overarching and cross-cutting criteria.

11 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.

12 The consultants carrying out the assessments of quality of evaluation reports and terms of references are referred to as 'raters'.

13 Organisation for Economic Co-operation and Development: Development Assistance Committee (OECD DAC).

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

15 Terms of Reference, Appendix 1: Guidance Manual: Quality Assessment Manual for Decentralised Evaluations and Reviews.



The scoring tool uses a four-point scale, as illustrated in Table 1. The Guidance Manual also provides a general description of scores. For evaluation reports, the Guidance Manual presents a qualifying statement and detailed scoring guidance for each quality criterion. For terms of references, no detailed scoring guidance is provided for individual quality criteria. Therefore,

the general description of scores was used as scoring guidance for all terms of reference quality criteria.

The templates for quality assessment of decentralised evaluations and terms of references were slightly modified in agreement with the Evaluation Department and converted to more user-friendly scoring protocol

formats.¹⁶ Space was added to allow for inclusion of administrative data, key findings presented in the reports and the rater's general comments. Annex 2 presents the final versions of the scoring templates. During the piloting and calibration process, the Guidance Manual was thoroughly discussed and where needed, complemented with clarifying comments.

Table 1: **Scoring Scale Used in the Quality Assessment Process**

Satisfactory Quality criteria are met to a good or adequate level		Less than satisfactory Quality criteria are not met to an adequate level	
4	Good quality The evaluation covers all the specified requirements or there are no substantial shortcomings in relation to the quality statement	2	Less than adequate quality The evaluation contains some elements of good application of the quality requirements
3	Adequate quality There are some shortcomings but the overall quality is still satisfactory	1	Poor quality The quality criterion was not applied or the quality of what was delivered when applying the quality criterion did not meet the requirements of the quality criterion
N/A	Not relevant/not applicable The quality criterion is not relevant/was not included in the evaluation		

Source: *Guidance Manual: Quality assessment manual for reviews and decentralised evaluations*, pp. 2–3.

¹⁶ The evaluation criterion coherence and the cross-cutting issue of human rights were added.



2.2 Identification and Selection of Decentralised Evaluations

Reports and terms of references were retrieved from *Evalueringssportalen* and *norad.no* and through email requests to Norad, the Ministry of Foreign Affairs and embassies. The emails specified the type of reviews and decentralised evaluations that the assessment team was interested in and requested respondents to send or share evaluation reports and accompanying terms of references for the period 2017–19.¹⁷ The assessment team sent a total of 81 emails and the Ministry of Foreign Affairs sent follow-up emails. A total of 26 embassies, 5 Norad departments and 2 departments of the Ministry of Foreign Affairs responded.¹⁸ A total of 129 evaluation reports were

¹⁷ See Annex 5 for a copy of the email and a list of respondents. The requests covered a three-year period to ensure that a sufficient number of reports and terms of references was received. It was later agreed that 2017 reports were not needed.

¹⁸ As the Ministry of Foreign Affairs forwarded some responses and some embassies are in charge of more than one country, the number of respondents and responses is not exact. Replies were received from embassies in the following cities and countries: Abuja, Accra, Amman, Beijing, Belgrade, Bogota, Brazil, Chile, Colombo, Cuba, Dar es Salaam, Delhi, Dhaka, Haiti, Havana, Juba, Kampala, Khartoum, Lilongwe, Luanda, Malawi, Rio de Janeiro, Tanzania, Uganda and Yangon. Norad departments that responded were the Health Section, Department for Education and Global Health, Education Section, Department for Climate change and Environment and the Knowledge Bank. Ministry of Foreign Affairs respondents were the Department for UN [United Nations] and Humanitarian

received via email and retrieved online.¹⁹ Excluding reports that were incomplete, were duplicates or did not meet the criteria discussed in Chapter 1,

a total of 55 decentralised evaluations implemented in 2018 or 2019 remained (Table 2).²⁰

Table 2: Number of Decentralised Evaluations Included in the Quality Assessment

Year	Total # of evaluations	Total # of terms of references	Evaluations without terms of references
2018	24	18	25%
2019	31	21	32%
Total	55	39	29%

Note: A list of reports with additional information is available in Annex 6.

Affairs, Department for Sustainable Development, and Section for South Asia and Afghanistan.

¹⁹ This included 32 reports produced in 2017, 39 produced in 2018, 48 produced in 2019 and some other documents (for example, evaluation briefs).

²⁰ See Annex 6 for a list of reports and distribution of reports across commissioner, implementer, thematic area and country.

2.3 Scoring Process and Data Analysis

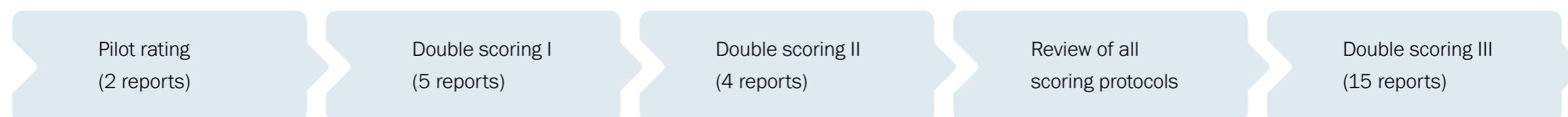
All team members participated in the quality assessment process, making it possible to draw upon a rich pool of experience from various geographic and thematic areas of the aid sector. Team members' experience and competences were matched to the theme and context of the evaluated interventions. To reduce the sources of bias, the raters' CVs were checked and the team members were asked for

information about prior relationships (e.g. as employer or employee, colleague or friend) with evaluators or with the subject under evaluation. When such a relationship was identified, the distribution of reports was adapted. Finally, evaluation reports authored by the same firm or consultant were assigned, to the extent possible, to different raters.

The scoring process comprised several steps aimed at ensuring consistency across raters and production of high-quality data, as illustrated in Figure 1. Throughout

the process, comments and clarifications were documented and shared with the team to ensure that everyone had access to the same information and applied the scoring tool consistently. The team held one all-day and several shorter team meetings (of two to four hours) to discuss the scoring tool, the scoring process, the quality of decentralised evaluations and findings from the evaluation reports.

Figure 1: [The Scoring Process](#)



The scoring process started with pilot scoring of two reports. Each report was scored by all raters and the scores were discussed in detail to arrive at a joint understanding of the scoring tool. To further calibrate the scoring, the first five decentralised evaluation reports were scored by two raters each (double scoring), reviewed by the quality assurer and revised. Four additional reports were double scored during the scoring process. After all reports and terms of references were scored, both the quality assurer and the team leader commented on the scores and justification comments. The raters then revised their scores or clarified their justification comments. Based on the number of comments by the quality assurer, the team leader selected and double-scored an additional 15 decentralised evaluations.

Once the scoring process was completed, the scoring protocols were transferred to Excel spreadsheets for analysis. The analysis process consisted of the following:

- quantitative analysis using various Excel features including descriptive statistics of the quality of evaluation reports and terms of references, comparison of scores within and between quality

areas and other variables, and comparison of scores across raters to identify remaining bias

- qualitative analysis through a review of justification comments for each quality criterion to assess the reliability of scores and collect examples for the findings chapter of this report
- discussions within the assessment team including discussions guided by results from the quantitative analysis and scoping discussions to capture individual perceptions of the scoring tool and process
- analysis and summary of the three best evaluation reports to illustrate examples of best practise (see Annex 4).

2.4 Limitations

The team assessed the quality of a sample of evaluation reports. The sample consists of evaluation reports and terms of references that were voluntarily sent to the team or uploaded to *norad.no* and *Evalueringssportalen*. No information was available

about the total number of decentralised evaluations that were produced during the period. Hence, the representativity of the sample is not known and the findings presented in this report cannot be generalised to other decentralised evaluations. A possible sampling error is that reports of poorer quality were not shared to the same extent as reports of higher quality. If this is the case, the findings in this report present an overly positive picture of the quality of decentralised evaluations.

Reliability of data in this assignment refers to the extent to which the quality assessment tool is consistently applied. The main sources of bias – that is, differences in the assessed quality that are not motivated by differences in the actual quality – are variations between raters and between assessments by the same rater (for example, over time) and bias relating to the evaluator, commissioner, type of intervention and/or context. As noted, extensive measures were taken to reduce bias. Nevertheless, some sources of bias remain. These relate to quality criteria that require subjective assessments. Examples are ‘methodological appropriateness’ and ‘feasibility of the terms of references’ and quality criteria where the scoring guidance is incremental in nature. In such criteria, the



scoring guidance lists the same items for all scores, with scores to be differentiated by whether, for example, there are some, many or no gaps in information. The definition of “some” and “many” is left to the raters to assess.

Validity refers to the extent to which the data collection tools measure what they are intended to measure. In this assignment, validity depends on whether the scoring tool, if applied correctly, gives an accurate picture of the quality of evaluation reports and terms of references and whether the evaluation reports and terms of references give an accurate picture of decentralised evaluations. The validity of the latter is difficult to determine without access to tenders, inception reports or management responses.

References to inception meetings and reports suggest to the raters that the evaluation process was often more thorough than the evaluation reports describe. In these cases, the quality of the evaluation would be higher than the quality of the evaluation report.

The extent to which the tool actually measures the quality of the evaluation reports and terms of references is lower for quality criteria where the detailed scoring guidance is additive in nature. In such quality criteria, new elements are added that must be included as the score increases. This implicitly assumes that elements are added in a specific order. When reports do not follow this order, strict application of the scoring tool does not give accurate assessments of quality.²¹ Additionally, for a few quality criteria, the scoring guidance omits important aspects of the quality criteria.²²

21 The quality statement for the quality criterion of ‘programme logic’ is an example of additive scoring guidance. The criterion definition includes that the programme logic is assessed in a comprehensive manner and that any gaps are identified; that the programme logic is assessed against relevant literature and/or evidence; and that assumptions underlying the programme logic are described. Score 2 requires that reference be made to relevant literature, while description of assumptions is required only for Score 4. In reality, many of the evaluation reports described assumptions made, but very few reports referred to relevant literature. For this particular quality criterion, the raters did not strictly follow the order of addition prescribed by the scoring guidance.

22 The definition of the quality criterion ‘efficiency’, for example, is that “[t]he report correctly interprets and assesses efficiency. It judges if the least costly resources possible are used in order to achieve the desired outputs. It may consider also whether alternative approaches would have produced the same results for less resources.” However, the scoring guidance completely focuses on the economy of the inputs against the quality of the outputs. Thus, the definition in the scoring guidance omits the other aspects of the criterion.

Overall, it is assessed that with the measures taken, reliability and validity are sufficiently high for the data to provide a good description of the quality of the decentralised evaluations that have been quality assessed.²³

2.5 Ethical Considerations

The main ethical consideration is related to potential conflicts of interest of the raters. The assessment team also recognises that the quality of evaluations is an important marketing asset for evaluators, and thus it is also important to consider the integrity of the raters. For these reasons, the team has taken great care in deciding who rates which evaluation and in keeping this information anonymous. The raters were clearly informed that the results would be presented in a way that does not reflect negatively on any specific evaluation report or terms of reference and that the identities of the raters of individual reports and terms of references would not be revealed. Apart from the three reports presented as good examples, all references to individual reports or terms of references are anonymous.

23 A comprehensive table of risks and limitations is presented in Annex 5.



Findings: The Quality of Evaluation Reports and Terms of References

The raters had access to terms of references for 39 of the 55 decentralised evaluations. For these, there is a positive but not very strong correlation between the quality of reports and quality of terms of references.

This chapter presents findings from the quality assessment of evaluation reports and terms of references, hereinafter referred to as decentralised evaluations. It begins with an overview, then presents the scores in different quality areas for evaluation reports and terms of references and highlights strengths and weaknesses. Additional figures and average scores for reports and terms of references are provided in Annex 3.

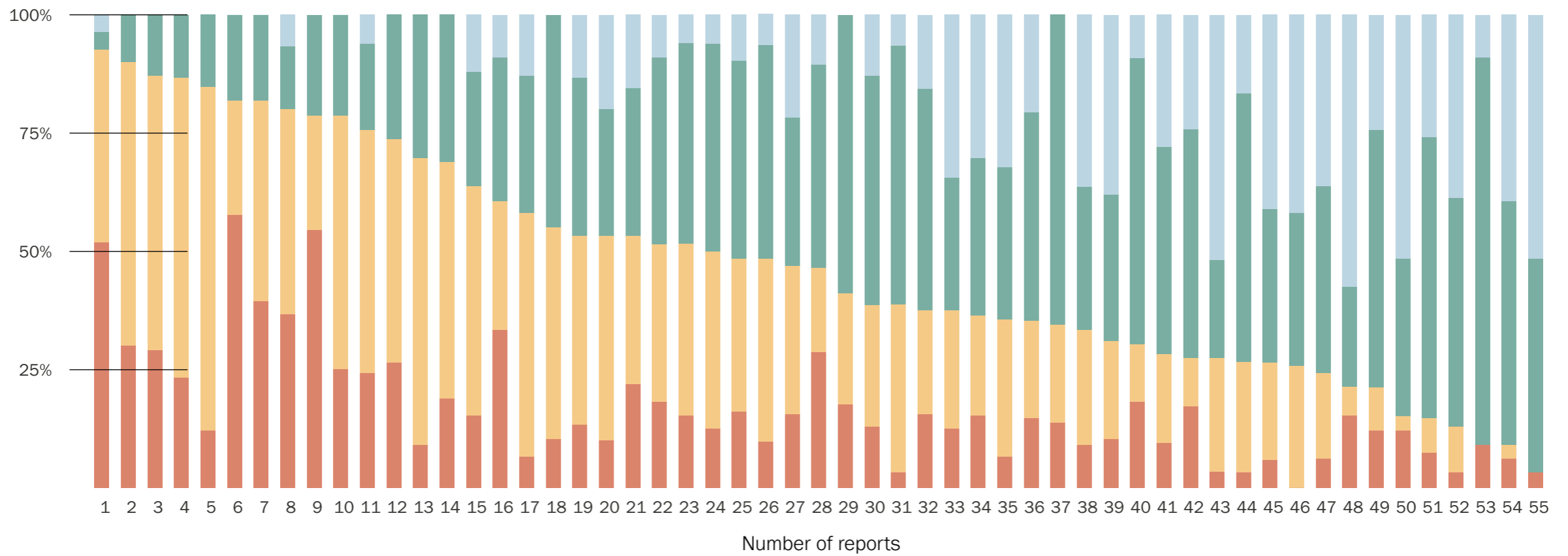
3.1 Findings and Distribution of Scores by Report

Figure 2 (next page) illustrates the distribution of scores for the 55 evaluation reports that raters assessed for quality. Using the colour coded scoring scale in

Table 1 in Chapter 2, poor quality (score 1) is indicated in red; less than adequate quality (score 2) in yellow; adequate quality (score 3) in green and good quality (score 4) in blue. This figure does not include the score of 'not applicable'. The reports are sorted so that the percentage of high scores (scores 3 and 4 shown in green and blue) increases as the figure is read from left to right. The first report starting from the left, received scores of 1 and 2 on more than 90% of the quality criteria; the last report on the right received scores of 3 and 4 on more than 90% of the quality criteria.²⁴

²⁴ Figure 12 in Annex 3 illustrates the distribution of scores across reports without sorting, and with identification numbers linking them to their respective terms of reference.

Figure 2: Distribution of Scores for Each Evaluation Report



Score 1 (red) Score 3 (green)
 Score 2 (yellow) Score 4 (blue)

Note: Each column represents the scores of one report. The larger the green and blue areas, the higher the quality of the report. The reports are sorted to group them in ascending (left to right) order of overall quality. Thus, report number 1 on the far left received scores of 1 (red) and scores of 2 (yellow), on over 90% of the quality criteria. At the other extreme, report number 55 received scores of 4 (blue) on 52% of the quality criterion, 3 on 45% of the quality criteria, and a score of 1 (red) on only one quality criterion. Score n/a is not included in this figure.

The figure illustrates clearly that the quality of decentralised evaluations needs improvement. Reports 1–24, or nearly half of the decentralised evaluations assessed, are rated as of poor or less than adequate quality (scores 1 or 2) on at least half of the quality criteria. On the other end of the spectrum, only 9 decentralised evaluations (reports 47–55) are rated as adequate or good quality (scores 3 or 4) on at least three-quarters of the quality criteria. None of the 55 reports are assessed to be of good or adequate quality on all quality criteria.

Some reports score consistently high or low across most quality criteria, while others have a large spread in quality across different quality criteria and areas. For example, the raters encountered several reports that were well-written, easy to read and nicely structured (i.e. that scored high on the first quality area) but had serious shortcomings, for example in the description of methodology or referencing to sources. Such reports pose a challenge to potential users as they may give the impression of being trustworthy and of high quality. The raters also encountered a number of reports with less perfect language but good descriptions of methodology and clear, evidence-based findings. Common to nearly all reports, though, is that the quality area with the lowest average score is methodology.

As noted in Chapter 1, the same quality is expected in reviews and decentralised evaluations. However, it was clearly expressed in some reports that because they were reviews rather than evaluations, the reader should expect a lower quality. A comparison of average scores on quality criteria for reports referred to as reviews (38 reports) and those termed evaluations (17 reports) confirms that evaluation reports are, on average, of higher quality than review reports.²⁵

The raters had access to terms of references for 39 of the 55 decentralised evaluations. For these, there is a positive but not very strong correlation between the quality of reports and quality of terms of references. The average of quality criteria scores on reports and terms of references have a correlation coefficient of 0.44, suggesting that the quality of reports is, to some extent, affected by the quality of terms of references. The raters also made an overall assessment of the quality of terms of references that is more strongly correlated (a correlation coefficient of 0.63) with the average quality of reports.

²⁵ See Figure 16 in Annex 3. The correlation coefficient between average scores on quality criteria of 'evaluations' and 'reviews' is 0.73.

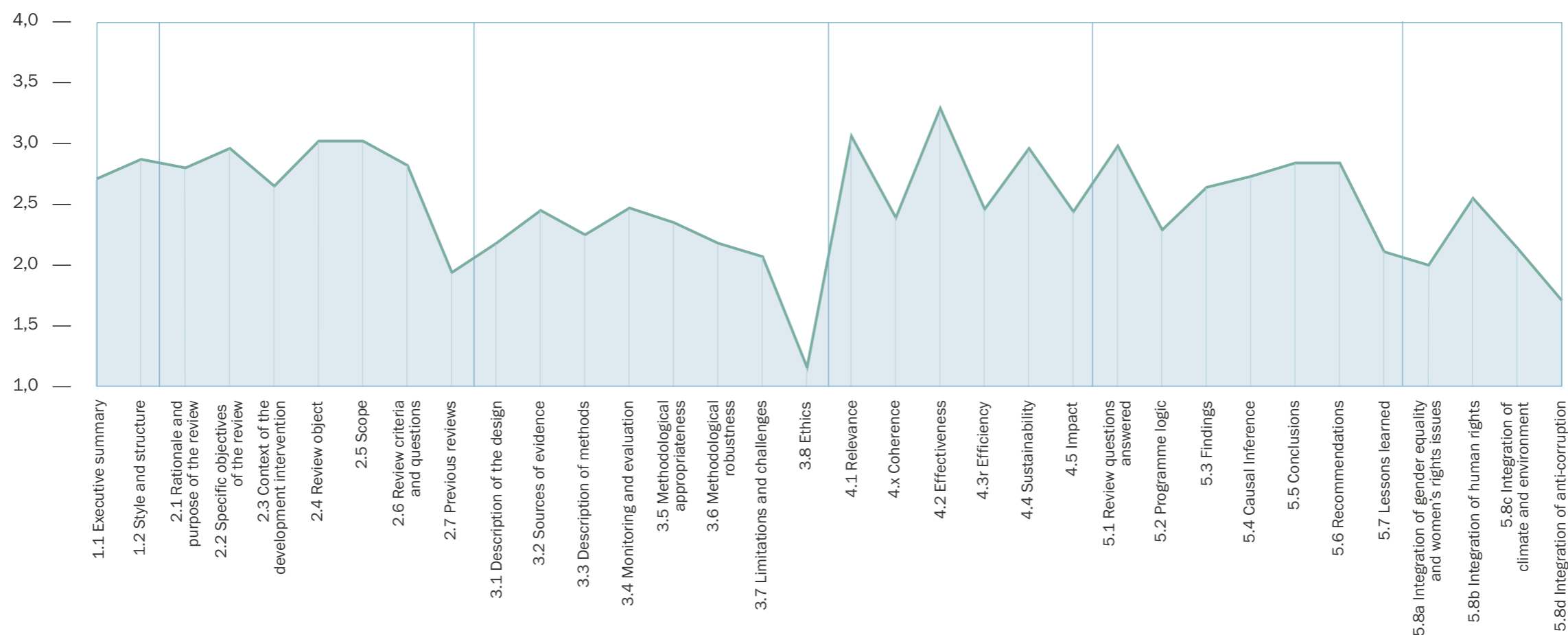
Annex 6 provides information about the 55 decentralised evaluations that were assessed. It is notable that several reports do not identify who conducted the evaluation: 13 reports do not name an author or evaluator and in 5 reports, it is not clear which organisation that conducted the evaluation.



3.2 Report Quality Criteria Averages

Figure 3 presents the average score for each quality criterion across all rated reports.

Figure 3: Average Scores: Report Quality Criteria



Note: Each segment separated by vertical blue lines covers one of the five quality areas. The green curve marks the average score for each quality criterion within the quality areas.

The average scores for quality areas 1 and 2 are between 2.5 and 3, shown in the first two segments of the figure. These scores relate to how well the report is written and how well it describes the assignment and background. Use of previous evaluations was rare and this criteria has a lower average score. The average of all scores in quality area 1 is 2.79 and the average of all scores in quality area 2 is 2.74.²⁶ The third segment of the figure shows average scores for criteria in quality area 3, which relates to the methodology used in the evaluation. Here, quality is markedly lower than for the first two quality areas, with the average of all scores only 2.14. While the average score for most quality criteria is between 2 and 2.5, the average score for ethics, an issue that very few reports even mentioned, is only 1.16. These results are in line with the findings of Evaluation Department (2017).

Quality area 4, shown in the fourth segment of the figure, refers to the application of OECD DAC²⁷

evaluation criteria.²⁸ While the average of all scores is relatively high (2.83), the jagged shape of the curve illustrates the uneven application of the different OECD DAC evaluation criteria. The application of relevance, effectiveness and sustainability was of higher quality, with average scores between 3 and 3.3, while the application of coherence, efficiency and impact was of lower quality, with average scores around 2.4. Quality criteria 5.1 to 5.7, in the fifth segment, include the link from data, analysis and findings to conclusions, lessons learned, and recommendations. The average of all scores on these quality criteria is 2.66. The average score for individual quality criteria varies: quality was higher for criteria relating to conclusions, recommendations and whether evaluation questions were answered; it was lower for presentations of programme logic and lessons learned. The sixth segment of the figure on the far right covers cross-cutting issues.²⁹ The average of all scores on cross-cutting issues is low (2.12), with criteria

averages ranging from 1.71 for anti-corruption to 2.55 for integration of gender equality and women's rights.

It is notable that only four quality criteria have an average score of 3 or higher, and then just barely. For the vast majority of quality criteria, the average score was between 2 and 3 and for half of the criteria, the average score was below 2.5. Moreover, none of the quality criteria relating to methodology have an average score above 2.5. The average score was below 2 for three quality criteria: the use of previous evaluation findings, the integration of anti-corruption, and description of ethical issues, which has an average score of only 1.16.

These findings again are in line with the results of Evaluation Department (2017), which found the highest quality to be in areas related to stating the purpose of the evaluation, defining the object to be evaluated, answering the questions posed in the terms of reference and making useful recommendations. The areas with the lowest scores in Evaluation Department (2017) were in areas such as describing the methods to be used in the review, dealing with ethical issues and examining the programme's logic.³⁰

²⁶ These scores are calculated as the average of the scores of all reports for all quality criteria within the quality area.

²⁷ Organisation for Economic Co-operation and Development: Development Assistance Committee ((OECD DAC).

²⁸ These include the six international evaluation criteria defined in OECD DAC (2019), Better Criteria for Better Evaluation Revised Evaluation Criteria Definitions and Principles for Use, OECD DAC Network on Development Evaluation.

²⁹ The cross-cutting issues within Norwegian aid are human rights, gender equality and women's rights, climate and environmental issues, and anti-corruption. These issues are to be considered at all levels and aspects of aid programming.

³⁰ See page 24 of Evaluation Department (2017).



3.3 Report Quality Criteria Scores

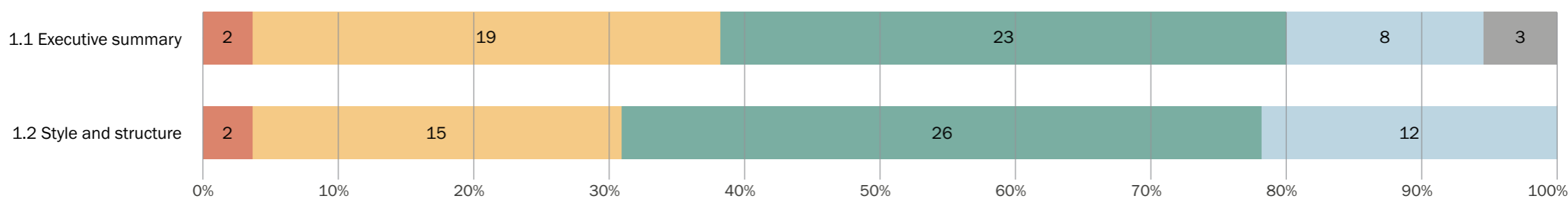
This section presents the scores for each report quality criterion in greater detail, together with examples and information drawn from the raters' justification comments. In Evaluation Department (2017), the scores are presented in two categories only. Scores 1 and 2 are aggregated into the 'less than adequate quality' category and scores 3 and 4 are aggregated into the 'adequate quality' category.

This report presents all scores. This gives a more nuanced picture of the results and avoids the incorrect impression that the difference between scores 2 and 3 is more distinct than between scores 1 and 2 or scores 3 and 4. For the purposes of comparability between the 2017 review and this report, shares for the aggregate categories are presented in the text. Figure 14 in Annex 3 presents scores for all quality criteria; below, one quality area at a time is analysed.

3.3.1 QUALITY AREA 1: SUMMARY, STYLE AND STRUCTURE

This quality area pertains to the content and presentation of the executive summary and the style and structure of a decentralised evaluation report. These quality criteria affect the overall impression and user-friendliness of these reports (Figure 4).

Figure 4: **Distribution of Scores: Summary, Style and Structure**



Score 1 Score 3 Score n/a
Score 2 Score 4

Criterion	1.1	1.2
Mean	2.71	2.87
SD	0.7	0.79

Note: The colours of the bars correspond to the quality scoring scale presented in Table 1 and the numbers are the number of reports that received the respective score. Orange is score 1 (poor quality), yellow is score 2 (less than adequate quality), green is score 3 (adequate quality), blue is score 4 (good quality) and grey is not relevant or not applicable. The table to the left shows the mean and standard deviation (SD) for each quality criterion for the assessed reports.

The most common shortcoming was that the executive summary was missing information about methodology, rationale, purpose and objectives. Some executive summaries did not present findings and several included an unabbreviated list of recommendations.

The average score for the 'executive summary' criterion is 2.71. In most evaluation reports, the executive summary gives a good summary of the report, with some or minor gaps in information; 23 reports are rated as being of adequate quality and 8 as of good quality. Four reports had no executive summary and received a 'not applicable' score. In the remaining 21 reports, the executive summary was incomplete and had many gaps in information. The most common shortcoming was that the executive summary was missing information about methodology, rationale, purpose and objectives. Some executive summaries did not present findings and several included an unabbreviated list of recommendations. The length of executive summaries ranged from less than one page to more than ten pages, and a number of reports ignored instructions in the terms of reference regarding the maximum length of the executive summary.

The average score on 'style and structure' is 2.87. Nearly 70% of the reports were assessed to be of adequate or good quality (score 3 or 4) and more than 20% were of good quality (score 4). Problems noted by the raters included overly long and dense reports that were difficult to read and poorly structured reports in which it was difficult to follow the line of evidence from data to findings and conclusions. The same lack of clarity regarding the presentation of findings, conclusions and recommendations was noted in Evaluation Department (2017).³¹

As was the case with executive summaries, several reports exceeded the page limit stated in the terms of references – some substantially so. Other reports referred readers to annexes for some information. Doing so was often a good way to adhere to the required page limit and make the report accessible, but it also had the effect in some reports of making their structure less logical because the background and contextual information, for example, were presented in annexes.

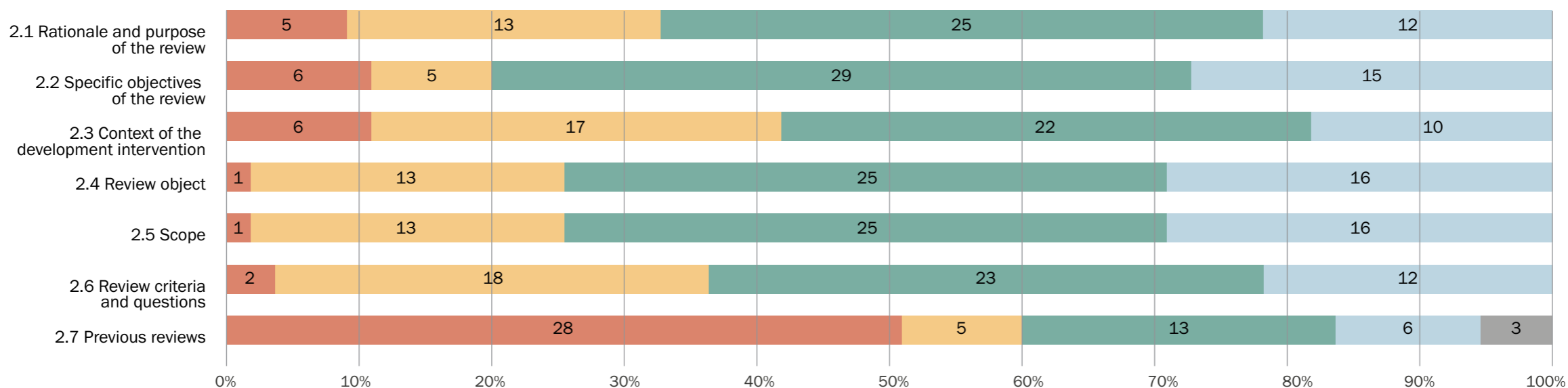
31 Evaluation Department (2017), p. 25.



3.3.2 QUALITY AREA 2: EVALUATION PURPOSE, OBJECTIVES AND SCOPE

This quality area includes criteria that describe the assignment and object under evaluation. Figure 5 shows the distribution of scores in this quality area.

Figure 5: Distribution of Scores: Evaluation Purpose, Objectives and Scope



Score 1 (Red), Score 2 (Yellow), Score 3 (Green), Score 4 (Blue), Score n/a (Grey)

Criterion	2.1	2.2	2.3	2.4	2.5	2.6	2.7
Mean	2.80	2.96	2.65	3.02	3.02	2.82	1.94
SD	0.88	0.89	0.90	0.77	0.77	0.81	1.12

Note: The colours of the bars correspond to the quality scoring scale presented in Table 1 and the numbers are the number of reports that received the respective score. The table to the left shows the mean and standard deviation (SD) for each quality criterion.

Description of the Assignment

Quality criterion 2.1 asks if the rationale, purpose, intended users and intended use of the evaluation are stated clearly. A third of the reports were rated as being of poor or less than adequate quality (scores 1 or 2). The most frequent shortcomings relate to the rationale, i.e. why the evaluation was undertaken, and the intended users. In the Guidance Manual the quality criterion ‘review object’ refers to the intervention that is being evaluated and ‘scope’ to the specific parts of the intervention that are included in the evaluation. In a majority of the reports assessed, the scope of the evaluation included the entire intervention and an explicit description of the scope was often missing. There were several examples of misplaced information and confusion regarding terminology: in several reports, the term ‘scope’ was used to describe the questions the report should answer and information about purpose and specific objectives was often placed under the wrong heading.³²

³² During the scoring process, the raters focused on the information presented, irrespective of where in the report it was presented.

Three criteria – specific objectives, evaluation object and scope – had some of the highest average scores, at 2.96, 3.02 and 3.02, respectively.³³ The criterion of ‘review criteria and questions’ was rated slightly lower, with 38% of the reports being of poor or less than adequate quality (scores 1 or 2). The main reason for a low score was the absence of a description of evaluation questions or cross-cutting issues, although some reports presented neither evaluation criteria nor specific evaluation questions. Apart from making it difficult to know what the evaluation intended to discover, a lack of these also made it difficult for the raters to assess quality on other quality criteria.

In many evaluation reports, the information referring to the description of the evaluation assignment was more or less a copy of the language in the terms of reference. For this reason, the raters do not consider

³³ Note that the high score for quality criteria 2.5 (scope) is due partly to the team’s decision to score information irrespective of the terms used in reports. If it was evident that the report considered the entire intervention to be the scope, the raters took the description of the evaluation object into consideration when determining the score for the quality criteria of ‘scope’.

the relatively high scores on these criteria to indicate a main strength. Rather, this information is available and easy to access and should be included in all reports without exception. Furthermore, the scores reflect the quality of the terms of references as much as the quality of reports.

Context of the Development Intervention

The average score on the description of context was relatively low (2.65), and a large share of reports (44%) were of poor or less than adequate quality on this criterion. This partly reflects the Guidance Manual requirement that a broad range of topics be covered, including context relating to cross-cutting issues and donor policies that few reports mentioned. The length and content of context sections varied considerably. Some reports have lengthy descriptions of certain contextual aspects but miss other important aspects of the context; other reports provide just a few pages of concise and relevant contextual information. A number of reports focus on the history of cooperation and do not describe context relating to stakeholders or national policies, among others.



Previous Reviews

This criterion concerns whether findings and recommendations from previous evaluations are described and referred to. Three reports explicitly stated that there were no previous reviews or evaluations of the evaluation object and thus were rated as 'not applicable'. Another 29 reports received a score of 1, as they did not mention previous evaluation findings at all, making it impossible for the reader to know whether previous evaluations were ignored or did not exist. All remaining reports stated that there were previous evaluations, but only 19 reports referred to previous findings and recommendations in the text and linked previous

findings and recommendations to specific evaluation questions or objectives. While this quality criterion had one of the lowest average scores (1.94), the distribution of scores indicate that where previous evaluations are mentioned, their findings are also to some extent used. The main weakness lies in those evaluations of previously evaluated interventions that do not refer to or use these and where valuable information and opportunities for learning are thus lost.

The distribution of scores indicate that where previous evaluations are mentioned, their findings are also to some extent used. The main weakness lies in those evaluations of previously evaluated interventions that do not refer to or use these and where valuable information and opportunities for learning are thus lost.



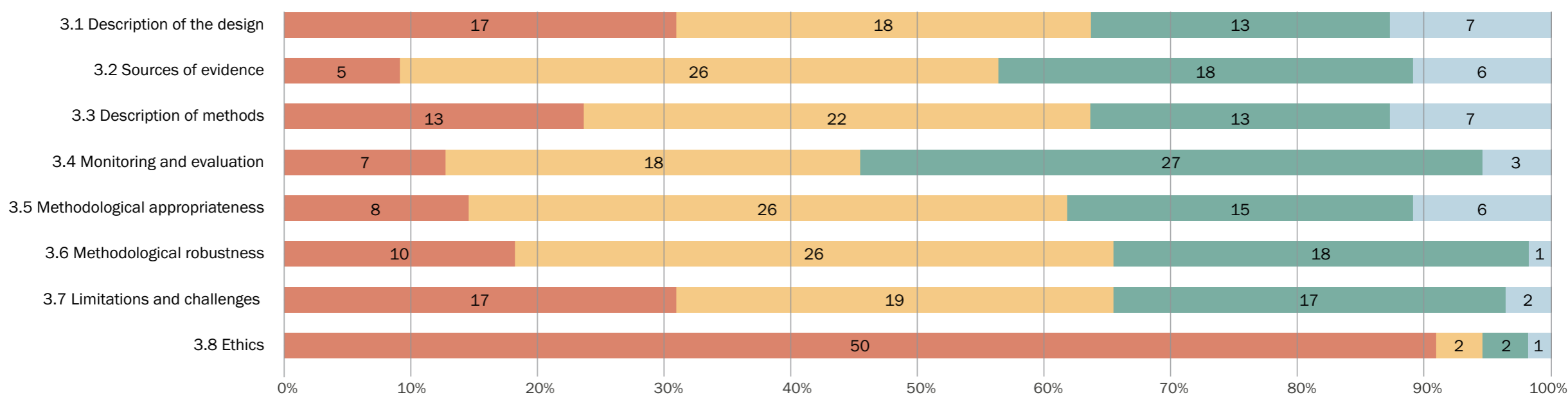
3.3.3 QUALITY AREA 3: METHODOLOGY

This quality area concerns how well various aspects of approach and methodology are described, whether monitoring and evaluation data are used, and whether ethical issues are considered (Figure 6). It should be

noted that the raters did not have access to inception reports, and thus the scores are based solely on the information presented in evaluation reports. Even so, a good description of methodology is important to the reader’s understanding and interpretation of

the evaluation results and an important part of an evaluation report. Several weaknesses were noted in this quality area and many of them are identical to those identified in Evaluation Department (2017).

Figure 6: Distribution of Scores: Methodology



Score 1 (Red), Score 2 (Yellow), Score 3 (Green), Score 4 (Blue), Score n/a (Grey)

Criterion	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8
Mean	2.18	2.45	2.25	2.47	2.35	2.18	2.07	1.16
SD	1.01	0.80	0.96	0.78	0.86	0.74	0.87	0.56

Note: The colours of the bars correspond to the quality scoring scale presented in Table 1 and the numbers are the number of reports that received the respective score. The table to the left shows the mean and standard deviation (SD) for each quality criterion.

The majority of the reports did not provide an adequate description of the approach, design and methods. This in turn implies that many decentralised evaluations do not provide the reader with enough information to fully understand how evidence, findings and conclusions were developed.

Approach, Design and Methods

The two quality criteria ‘description of the design’ and ‘description of methods’ refer to how well the report describes and justifies the way the evaluation is carried out. The first of these relates to the overarching approach or conceptual framework used for the evaluation. Examples of approaches include utilisation-focused evaluation, process evaluation and participatory evaluation. The Guidance Manual states that the design of an evaluation is how a conceptual framework has been operationalised to answer the evaluation questions. This description can include details about the evaluation process and how the conceptual framework will be implemented in relation to different components of the programme logic.

The scores for this quality criterion clearly indicate that this is a weakness in many decentralised evaluations. Almost a third of the reports assessed provided no information about a conceptual framework or overarching design; these received a score of 1. Another third included a brief overview of how the decentralised evaluation was implemented or a brief statement or name of an approach; these received a score of 2. Just over a third of the reports provided an adequate or good discussion (scores 3 or 4) of the conceptual framework and design, but only 7 of

these 20 reports explained why they had selected this approach.

The quality criterion ‘description of methods’ relates to how well the report describes instruments, techniques and methods for collecting and analysing data. Many reports stated that data were collected by document review, interviews and focus group discussions. Details about how this was done or how data were analysed were often not provided. Reports that failed to mention methods or only mentioned methods for collecting data were given a score of 1 or 2, depending on how clearly these methods were described. Only 20 reports also described the methods used for analysing data, (required for scores 3 and 4). A score of 4 requires very clear descriptions of data collection and analysis methods as well as a description of how gender-sensitive data were collected and analysed. Seven reports received score 4.

There is a relatively strong correlation between the criteria ‘description of design’ and ‘description of methods’.³⁴ Only 16 reports were assessed to be of adequate or good quality (scores of 3 or 4) on both

³⁴ The correlation coefficient between these two quality criteria is 0.74.



criteria. Eight reports lacked any information about both approach and design and methods; these received a score of 1 on both ‘description of design’ and ‘description of methods’. The ratings point to one of the most serious weaknesses of the assessed decentralised evaluations: the majority of the reports did not provide an adequate description of the approach, design and methods. This in turn implies that many decentralised evaluations do not provide the reader with enough information to fully understand how evidence, findings and conclusions were developed.

Sources of Evidence

In the scoring guidance, the criterion ‘sources of evidence’ considers whether no references to primary or secondary data are made (score 1), whether there is inconsistent referencing (score 2), or whether referencing exists and is consistent (scores 3 and 4). To achieve score 4, the report must also describe how the sources of evidence were selected. Notably, only six reports provided clear descriptions of sampling and selection strategies, i.e. they explained how it was decided whom to talk to, which organisations to visit, where to make field visits, etc. When this information is omitted, it is impossible to know whether the data

were collected in such a way that they represent a correct picture of reality or are biased, for example because only one type of stakeholder was interviewed.

The majority of reports (56%) made no reference or inconsistent references to primary and secondary data. A common feature was that background and context sections of the reports were relatively well-referenced and findings chapters contained fewer and less specific references to sources. Similarly, references to secondary sources were more common and more specific (i.e. they pointed to a specific document), while references to primary sources were less common and often vague to some degree. Examples of such language include “field notes show”, “respondents were of the view”, a group “thought” and such and such organisation “stated”. Details such as the share of respondents that had a particular view, their gender distribution and location, the type of interview, etc. were very rarely stated. Several reports combined extensive lists of interviewees or focus groups with a near absence of references to primary sources of information. These weaknesses are assessed to be serious, especially alongside the lack of information about data selection strategies. They make it difficult or impossible to assess the reliability and representativity

of the evidence that is used as basis for findings, conclusions and recommendations.

The quality criterion ‘monitoring and evaluation’ also refers to sources of information, but in the form of internal monitoring and evaluation data. The raters found that 7 reports made no use of existing monitoring and evaluation data (score 1); 18 reports used monitoring and evaluation data but without making any critical assessment of the quality of information; and 30 reports described strengths and weaknesses of the monitoring and evaluation data and then used the data in their analysis (score 3 or 4). Three of these 30 reports were assessed to have made excellent use of monitoring and evaluation data.

Methodological Appropriateness

This quality criterion refers to the appropriateness of the evaluation methodology (including approach, design, methods for data collection, analysis and sampling) given the evaluation purpose, objectives and questions and to whether the methodology is well-justified in the report. The criterion thus requires an assessment of both how well the report explains why a certain methodology was selected and how appropriate this methodology is. In the many reports that lacked good descriptions of approach



and methods, the raters had to rely on their experience as evaluators and make assessments based on information pieced together from other parts of the reports.

The score for ‘methodological appropriateness’ is highly correlated with ‘description of design’ and ‘description of methods’.³⁵ Eight reports completely lacked information justifying the choice of methodology or had an inappropriate methodology (score 1). In nearly half (26) of the reports, the methodology was assessed to be appropriate although the report did not provide a clear justification or link to evaluation questions. Just over a third of the evaluations were assessed to be of adequate or good quality (scores 3 or 4), but only six of these provided a good enough justification for ‘methodological appropriateness’ to receive the highest score.

Low quality on the criteria relating to approach and design, methods, sources of evidence, and methodological appropriateness has three main effects. First, if the approach and methods are not appropriate, the findings of the evaluation may not be correct.

³⁵ The correlation coefficient for ‘methodological approach’ and ‘description of design’ is 0.75, the correlation coefficient for ‘methodological approach’ and ‘description of methods’ is 0.73.

Second, it becomes difficult or impossible for the reader to understand how the evaluator has selected, collected and interpreted the evidence upon which findings and conclusions are based. Third, transparency of the evaluation process is limited, which not only makes it difficult to follow the line of evidence but also reduces the credibility of conclusions and recommendations. This is a serious weakness of the decentralised evaluations that were rated.

Methodological Robustness

This quality criterion refers to the extent to which the evidence and findings are credible. According to the Guidance Manual, evidence must be triangulated and reliability and validity of data must be assessed to score high on this criterion. The OECD glossary of key terms defines triangulation as the “use of three or more theories, sources or types of information, or types of analysis to verify and substantiate an assessment” and notes that “by combining multiple data sources, methods, analyses or theories, evaluators seek to overcome the bias that comes from single informants, single methods, single observer or single theory studies”.³⁶

³⁶ OECD (2002). Glossary of key terms in evaluation and results based management, p. 37.

First, if the approach and methods are not appropriate, the findings of the evaluation may not be correct. Second, it becomes difficult or impossible for the reader to understand how the evaluator has selected, collected and interpreted the evidence upon which findings and conclusions are based. Third, transparency of the evaluation process is limited, which not only makes it difficult to follow the line of evidence but also reduces the credibility of conclusions and recommendations.



In addition, 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”.³⁷

Triangulation thus implies an active comparison of data from different sources or methods. A long list of interviewees, for example, does not imply triangulation unless the information from different sources is actually used and compared. This can be indicated by statements such as “the annual report stated that 22 trainings were held; this was confirmed by all staff interviewed”. One report was assessed to be of good quality on this criterion. In 18 reports, evidence was relatively consistently triangulated and there was some discussion of reliability and validity of the data (score 3). Notably, 10 of the reviewed reports did not triangulate evidence (score 1) and another 26 reports

only triangulated some of the evidence (score 2). These 36 reports also lacked a discussion of the reliability and validity of data. This implies that in 65% of the reports assessed for quality, no or little comparison was made of evidence from different sources and that there was no critical analysis of the credibility of the data used as the basis for findings, conclusions and recommendations. This is especially problematic where there is also no or inconsistent referencing of sources of information, which was the case for over half of the reports.

The assessment team found serious shortcomings in both application and presentation of this quality criterion. Methodological robustness is thus a main weakness in the assessed decentralised evaluations as triangulation, reliability and validity are all crucial to ensure that an evaluation rests on solid evidence. The low scores on this criterion also signal an underlying problem of lack of attention to and transparency about the quality of evidence in decentralised evaluations.

Limitations and Challenges

Almost a third of the assessed reports did not describe limitations at all (score 1). Another third provided some information but often merely listed limitations relating to the implementation of the evaluation, such as lack of time, access to documents and logistics. The reports that received score 3 provided some discussion of limitations in relation to data sources, sampling or selection; data collection and analysis; and mitigation measures. Where implications of the limitations were mentioned, these often related to the number of interviews or field visits and rarely to validity and reliability of data. Score 4 requires that the evaluation report describe limitations related to the sample’s representativeness, how these affect the results of the evaluation, and any obstructions to a free and open review process. Only two reports fulfilled these requirements. The low scores on this quality criterion strengthens the impression that there is a serious weakness relating to analysing and describing the quality of data and methods used in decentralised evaluations.

³⁷ Ibid. p. 32.



Over 90% of the reports did not address ethical issues at all. The absence of any mention of ethical issues in the evaluation reports is a serious weakness, especially if it also reflects a lack of attention to ethical issues in the evaluation process.

Ethical Issues

The OECD DAC Quality Standards for Development Evaluation describe evaluation ethics as follows:

“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.”³⁸

The scoring guidance for this quality criterion refers to how well an evaluation report describes ethical issues and the approach taken to address them and whether these approaches are appropriate. Only 5 of the 55 decentralised evaluation reports described ethical issues, and 3 of the 5 also described the approach taken to address the ethical issues. This finding suggests that over 90% of the reports did not address ethical issues at all. The absence of any mention of ethical issues in the evaluation reports is a

³⁸ OECD (2010), p.6.

serious weakness, especially if it also reflects a lack of attention to ethical issues in the evaluation process.

3.3.4 QUALITY AREA 4: APPLICATION OF OECD DAC EVALUATION CRITERIA

This quality area assesses whether and how well an evaluation applies the OECD DAC evaluation criteria – henceforth referred to as “evaluation criteria” to be in line with the terminology of OECD DAC (2019). The terms of reference normally specify the evaluation criteria to be applied. If an evaluation criterion was neither asked for nor applied, a score of ‘not applicable’ was used. Score 1 was used when a report did not apply an evaluation criterion although the terms of reference requested it. Some reports applied evaluation criteria that were not requested by the terms of reference; this added to the breadth of the analysis but reduced focus on the intended evaluation criteria. Such reports were rated as to how well the evaluation criteria were applied. Hence, when an evaluation criterion was applied in the report, the raters scored what was written, whether or not this was requested in the terms of reference.³⁹ The

³⁹ As the assessment team did not have access to all terms of references, the team agreed this was the most appropriate approach. If the terms of reference were not available and the evaluation report neither listed the evaluation criterion among evaluation questions nor assessed it, ‘not applicable’ was assigned.

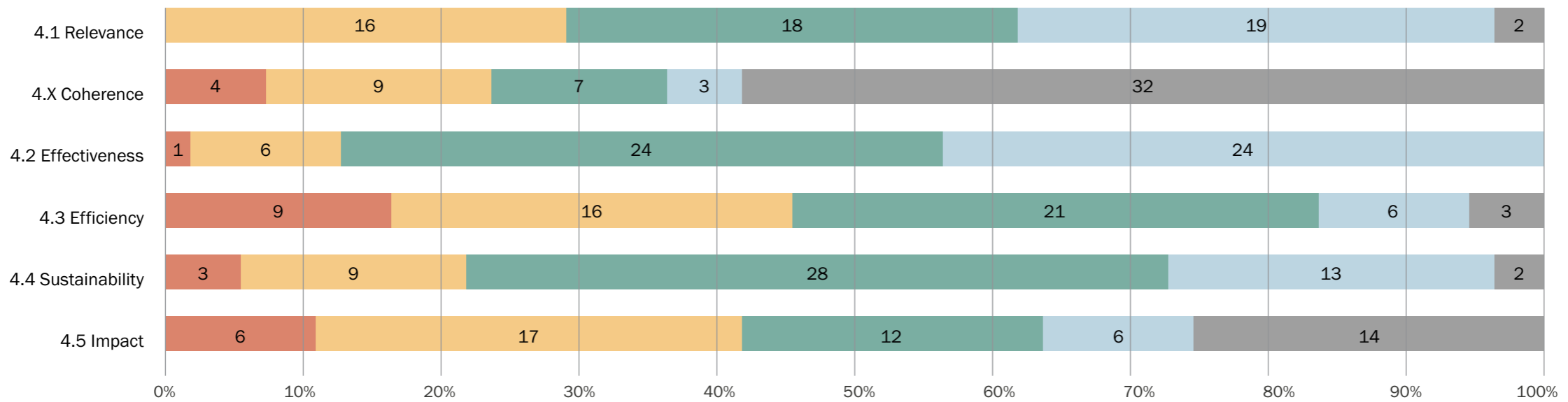


scoring guidance allows for many gaps in information for score 2 and some gaps for score 3. To achieve score 4, complete and detailed information about

the application of the evaluation criterion must be presented. Figure 7 shows that there is a large variation in this quality area, both in terms of the number of

reports in which the evaluation criterion was applied and in terms of quality.

Figure 7: Distribution of Scores: Application of Evaluation Criteria



Score 1 (Red), Score 2 (Orange), Score 3 (Green), Score 4 (Blue), Score n/a (Grey)

Criterion	4.1	4.x	4.2	4.3	4.4	4.5
Mean	3.06	2.39	3.29	2.46	2.96	2.44
SD	0.81	0.92	0.73	0.91	0.80	0.91

Note: The colours of the bars correspond to the quality scoring scale presented in Table 1 and the numbers are the number of reports that received the respective score. The table to the left shows the mean and standard deviation (SD) for each quality criterion.

Relevance and Coherence

Both relevance and coherence assess how well-suited an intervention is to its context. Relevance refers to donors, recipients and target group; coherence refers to other ongoing interventions. Relevance stands out as the only quality criterion for which no report received a score of 1. The Guidance Manual requires that relevance be assessed in relation to the priorities and policies of both target group, recipient and donor, and a common shortcoming was the omission of relevance in relation to donor policies. A less frequent shortcoming was not discussing relevance to the target group.⁴⁰ However, all reports made at least some assessment of relevance and two-thirds of the reports were assessed to be of adequate or good quality (scores 3 or 4).⁴¹

Coherence assesses the compatibility of an intervention with other interventions in the country, sector or institutions. OECD DAC (2019) added it as

40 In some cases, the terms of references specified that certain aspects of relevance should be assessed. As the raters did not have access to terms of references for all evaluations, the scoring guidance was strictly followed and such information was not taken into consideration. Thus, some reports may fulfil all requirements relative to the terms of references even if they did not receive a high score.

41 The exceptions were two reports that were rated 'not applicable' because their terms of references did not ask for an assessment of relevance.

an evaluation criterion, which may explain why it was discussed in just over 40% of the reports and why the quality of the assessments in relation to this criterion was relatively low.⁴² The scoring template for terms of references does not distinguish between different DAC criteria, but the raters have the impression that coherence was discussed in evaluation reports more frequently than demanded in terms of references. The reason may be that the criterion is common in evaluations of humanitarian aid.⁴³

Effectiveness and Efficiency

Effectiveness and efficiency assess different aspects of goal achievement. While effectiveness assesses whether the intended objectives are achieved, efficiency assesses whether they were achieved with the least costly means. The understanding and application of these two evaluation criteria were of very different quality. 'Effectiveness' stands out in several ways: it is the only evaluation criterion that was included in all reports, it is the quality criterion with the highest

42 See the revised evaluation criteria presented in OECD DAC (2019). This quality criterion was not included in Evaluation Department (2017) and has been added to the Guidance Manual to enable future comparison.

43 See, e.g., Overseas Development Institute (2006). Evaluating Humanitarian Action Using the OECD-DAC Criteria: an ALNAP Guide for Humanitarian Agencies.

'Effectiveness' stands out in several ways: it is the only evaluation criterion that was included in all reports, it is the quality criterion with the highest average score (3.29) and it is the quality criterion with the largest number of reports that received a score of 4.



average score (3.29) and it is the quality criterion with the largest number of reports that received a score of 4. The Guidance Manual states that ‘effectiveness’ is the likelihood that the intervention will achieve its objectives. The raters assessed that 87% of the reports applied this criterion well enough to be of appropriate or good quality (score 3 or 4). In addition, 24 of these reports discussed risk factors and how these were managed and received a score of 4. A main strength of the decentralised evaluations rated is clearly that they assess effectiveness appropriately and describe this well.

For all quality criteria, the Guidance Manual first presents a quality statement and then presents scoring guidance for each score. The quality statement for ‘efficiency’ is as follows: “The report correctly interprets and assesses efficiency. It judges if the least costly resources possible are used in order to achieve the desired outputs. It may consider also whether alternatives approaches would have produced the same results for less resources”.⁴⁴ While this statement leaves room for a variety of approaches to assess efficiency, the scoring guidance refers only to whether the report assesses the economy of the

inputs against the quality of the outputs and to how detailed the account of inputs is in relation to outputs. Very few reports explicitly assessed the economy of inputs against the quality of outputs, but several reports assessed efficiency in other ways that were in line with the revised OECD DAC (2019) definition of the evaluation criterion.⁴⁵ Examples include discussions of efficiency in terms of organisational structure, human resources, and management and financial systems. In such cases, the raters relied more heavily on the quality statement than on the scoring guidance for ‘efficiency’.

Although efficiency does not have the lowest average score, it is one of the weakest areas in most evaluation reports. Only six reports demonstrated a correct interpretation and assessment of efficiency, made a thorough assessment of whether the least costly resources possible were used to achieve the desired outputs and considered alternative approaches to

⁴⁵ See OECD DAC (2019), p. 10. The revised definition reads: “Efficiency: how well are resources being used? The extent to which the intervention delivers, or is likely to deliver, results in an economic and timely way. Note: ‘Economic’ is the conversion of inputs (funds, expertise, natural resources, time, etc.) into outputs, outcomes and impacts, in the most cost-effective way possible, as compared to feasible alternatives in the context. ‘Timely’ delivery is within the intended timeframe, or a timeframe reasonably adjusted to the demands of the evolving context. This may include assessing operational efficiency (how well the intervention was managed).”

achieve the results. These six reports were assessed to be of good quality and received a score of 4. An additional 21 reports were assessed to be of adequate quality (score 3). Nearly half of the decentralised evaluations that were requested to assess efficiency did not provide an adequate description or analysis of the concept (score 2) or did not assess efficiency at all (score 1).

Nearly half of the decentralised evaluations that were requested to assess efficiency did not provide an adequate description or analysis of the concept (score 2) or did not assess efficiency at all (score 1).

⁴⁴ See the Guidance Manual, p. 13.



Examples include interpreting a low rate of expenditures versus budget as poor efficiency and listing expenditures without discussing if these were reasonable or linking them to results. Few reports compared costs with other similar initiatives or assessed whether alternative approaches could achieve the same result. Some reports clearly stated that they could not assess efficiency. Others were less transparent. This weakness is particularly serious as it implies that a large number of decentralised evaluations make recommendations, for example regarding continued funding, without consideration of the cost of the intervention relative to its results.

Sustainability and Impact

The quality criterion ‘sustainability’ relates to whether the report correctly interprets and describes sustainability and how well it describes the extent to which the benefits of the intervention are likely to continue after donor funding has been withdrawn. This evaluation criterion was applied in all but two reports. The average score for sustainability is 2.96, and only 12 reports were rated as being of poor or inadequate quality. Scores 3 and 4, awarded to 75% of the reports, require an evaluation report to clearly describe the likelihood that the benefits will continue after the intervention and to assess environmental and financial

sustainability. The 28 reports that received a score of 3 had some gaps in information; 13 reports received a score of 4 as they presented clear information in relation to all components of the intervention and, where applicable, made very clear assessments of environmental and financial sustainability.

The criterion ‘impact’ relates to whether the report correctly interprets and assesses impact, i.e. the extent to which the initiative is likely to or has begun to achieve its longer-term goals beyond the life of the intervention. The raters found that 14 reports did not discuss impact and were not required by their terms of references to do so; these were rated as ‘not applicable’. Six reports received a score of 1 either because they failed to assess impact when the terms of references demanded it or because they did not convey an understanding of the concept. Another 17 reports assessed impact in relation to what had happened or was likely to happen, but only discussed positive impact (score 2). The importance of identifying and discussing both positive and negative impact and looking for both intended and unintended impact was noted in 18 reports, of which 6 discussed this clearly enough to receive score 4. This implies that among the 35 evaluations that attempted to assess impact, only half did so with adequate or good quality

Among the 35 evaluations that attempted to assess impact, only half did so with adequate or good quality.

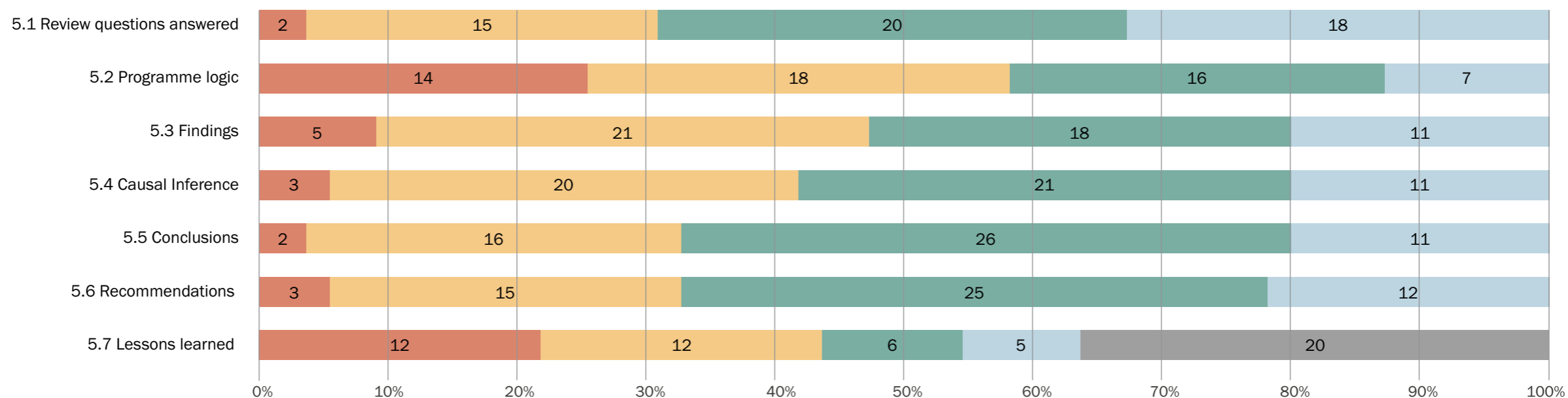


3.3.5 QUALITY AREA 5: ANALYSIS, FINDINGS, CONCLUSIONS, LESSONS LEARNED AND RECOMMENDATIONS

This quality area refers to analysis of evidence, traceability of conclusions and findings to data, linking of data and findings to programme logic, the basis for recommendations and lessons learned, and how well

these are expressed. There is a large spread in quality, both across reports and across criteria, with quality criteria averages ranging from among the highest to the lowest (Figure 8).

Figure 8: Distribution of Scores: Analysis, Findings etc.



Score 1 (Red), Score 2 (Yellow), Score 3 (Green), Score 4 (Blue), Score n/a (Grey)

Criterion	5.1	5.2	5.3	5.4	5.5	5.6	5.7
Mean	2.98	2.29	2.64	2.73	2.84	2.84	2.11
SD	0.86	0.98	0.90	0.84	0.78	0.83	1.04

Note: The colours of the bars correspond to the quality scoring scale presented in Table 1 and the numbers are the number of reports that received the respective score. The table to the left shows the mean and standard deviation (SD) for each quality criterion.

Evaluation Questions Answered

The scores on this criterion can be interpreted in two ways. On one hand, it has one of the highest average scores (2.98), and a large share of reports (69%) provide answers to the evaluation questions (scores of 3 and 4). Score 4 was awarded to 18 reports that fully answered the evaluation questions and also documented modifications, if any, to the questions that were presented in the terms of reference. On the other hand, 31% of the reports did not fully respond to the evaluation questions in the terms of reference.

While the structure of some reports followed the evaluation questions, helping both readers and the authors to remain focused on the issues identified in the terms of references, the structure of other reports used evaluation criteria as headings in the findings chapter. This latter option worked well in some cases, but in other reports the evaluation questions were lost in a more general treatment of the evaluation criteria. It is notable that only a third of the reports clearly documented whether changes were made to the evaluation questions set forth in the terms of reference. As such changes were not routinely stated in the evaluation reports, it is not possible to determine whether the lack of response to a specific evaluation question was in fact agreed to with the organisation

that commissioned the evaluation. This illustrates the risk that shortcomings in report writing may be interpreted as shortcomings in the evaluation process.

Programme Logic, Findings and Causal Inference

The quality criterion ‘programme logic’ refers to the extent to which the intervention’s theory of change or programme logic is assessed. A programme logic describes how a programme intends to reach its goal. Analysis of the programme logic helps evaluators understand how the programme works, how it is best evaluated and if it is on track to achieving its intended results. The raters allowed for a wide range of possible ways to describe how an intervention’s results were intended to be achieved, regardless of whether the evaluation referred to this as a programme or intervention logic, theory of change, or a results chain. Still, only 24 reports (42%) were found to be of adequate or good quality on this criterion, meaning that they appropriately assessed the programme logic and discussed gaps. Seven of these reports also described underlying assumptions (score 4). However, 25% of the reports did not describe the programme logic at all or did so very poorly (score 1).

The ‘findings’ quality criterion relates to whether findings are founded on evidence. To receive a score of 4, the

line of evidence should flow logically from the analysis, triangulation should be used appropriately and the gaps in data and the impact of such gaps on the findings must be discussed. While 11 reports did so, 18 other reports did not triangulate data although they showed a clear link to evidence (score 3). In 21 reports, there was a lack of clarity in the line of evidence for some findings and in the use of triangulation and discussion of data gaps. Five reports presented findings that were not appropriately founded on evidence (score 1). These scores suggest that almost half of the decentralised evaluations exhibited severe shortcomings in the line of evidence, triangulation, and presentation of data gaps and their consequences. These weaknesses are closely linked to the weaknesses identified in the quality criteria relating to appropriateness and robustness of methodology, and they diminish the transparency and credibility of the decentralised evaluations.

The criterion ‘causal inference’ refers to the extent to which the report differentiates between outputs, outcomes and impacts; demonstrates progression towards development results; and discusses attribution, contribution and influence of external factors. All but three reports made at least some distinction between outputs, outcomes and impact, showing an overall understanding of these concepts. A score of 2 was



attributed to 20 reports due to some shortcomings in terms of the distinction they drew between the three types of results and because they did not demonstrate progression towards the intended results. Another 32 reports (58%) were of adequate or good quality on this criterion in that they showed a clear understanding of the output, outcome and impact concepts, linked the discussion to development results, and discussed the extent to which the intervention was responsible for the results. Of these reports, 11 included discussion of attribution, contribution and external factors (score 4).

Conclusions and Recommendations

The distribution of scores is nearly identical for the criteria ‘conclusions’ and ‘recommendations’: the average score for both is 2.84 and two-thirds of the reports are of adequate or good quality. Very few reports presented conclusions or recommendations that were assessed to be not founded in evidence (score 1).

Not all reports made a clear distinction between findings and conclusions. Several presented a mix of findings and conclusions, often in a chapter dedicated to findings. Conclusions chapters were often very brief. In some cases, they had a clear focus and added value to the analysis, while in others they seemed to lack

purpose. Sixteen reports presented conclusions that, while drawing on evidence, were assessed as being not proportionate or reasonable given the strength of the evidence (score 2). The 37 reports that received scores of 3 or 4 presented conclusions that were both reasonable and pertinent to the purpose of the evaluation and 11 of these also identified priority issues (score 4).

In all but three reports, recommendations were assessed to be founded in evidence. The scoring guidance for scores 2, 3 and 4 sets out progressively higher demands for recommendations to be clear, relevant, targeted and actionable. While 12 reports fulfilled all these criteria (score 4), 15 reports presented recommendations that were based on evidence but were not clear, relevant, targeted and actionable (score 2). The way recommendations were presented varied widely, ranging from vague suggestions to very clear and targeted recommendations that were described in a way that was quite actionable.

Lessons Learned

OECD (2002) defines lessons learned as “generalizations based on evaluation experiences with projects, programs, or policies that abstract from the specific circumstances to broader situations.

Frequently, lessons highlight strengths or weaknesses in preparation, design, and implementation that affect performance, outcome, and impact.”⁴⁶ Twenty of the reports did not include lessons learned and twelve reports presented lessons learned that did not follow logically from the conclusions. Only eleven reports presented lessons learnt that followed logically from conclusions and also contributed to general knowledge (scores 3 and 4). Examples of misunderstandings include reports that summarise conclusions or discuss organisational learning under a heading of ‘lessons learned’.

Cross-cutting Issues

Norwegian development policy identifies four cross-cutting issues that are to be taken into consideration in all aspects of Norwegian development policy and aid. These are human rights, women’s rights and gender equality, climate change and environment, and anti-corruption. All development efforts are to be assessed on the basis of how they affect or are affected by these cross-cutting issues. It is natural to expect, then, that all decentralised evaluations would include assessments of these cross-cutting issues. Yet, this is not the case. The terms of references for a substantial number of

⁴⁶ OECD 2002, p.26.

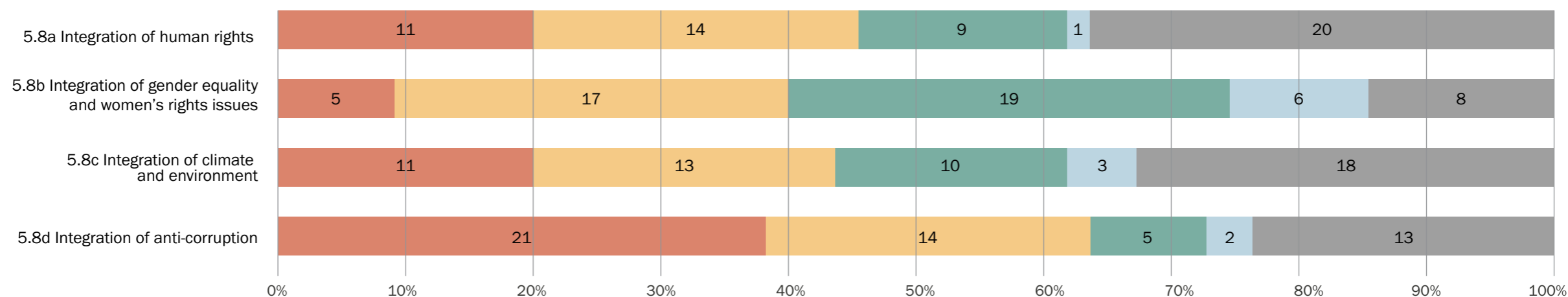


evaluations did not include one or more cross-cutting issues and are shown in Figure 9 as scores of ‘not applicable’ in the grey-coloured part of the bars. The red areas in the figure show that cross-cutting issues were

not always included even when the terms of references asked for this (score 1). In reports that did include cross-cutting issues, the level of integration varies greatly. While several reports made brief statements in response

to general evaluation questions on cross-cutting issues, only a few analysed them fully and also integrated them, for example by integrating gender aspects in evaluation design, data collection and analysis.

Figure 9: Distribution of Scores: Cross-cutting Issues



Score 1 (red), Score 2 (orange), Score 3 (green), Score 4 (light blue), Score n/a (grey)

Criterion	5.8a	5.8b	5.8c	5.8d
Mean	2.00	2.55	2.14	1.71
SD	0.83	0.85	0.93	0.85

Note: The colours of the bars correspond to the quality scoring scale presented in Table 1 and the numbers are the number of reports that received the respective score. The table to the left shows the mean and standard deviation (SD) for each quality criterion.

The scoring guidance defines the scores for all cross-cutting issues as follows: ‘not applicable’ means the issue was not asked for; score 1 means it was asked for but not addressed; score 2 means the issue was addressed, but only partially and with many gaps; score 3 means it was addressed with only some gaps; and score 4 means the issue was fully integrated in findings, conclusions and recommendations as appropriate.

The addition of human rights as a cross-cutting issue in Norwegian development policy is relatively recent, which may explain why so few reports received high scores on this criterion and why so many decentralised evaluations ignored it altogether. In 20 of the decentralised evaluations assessed, the terms of references did not include the issue of human rights; another 11 reports did not address human rights although the terms of references requested it. Many reports simply stated that the intervention being evaluated had no effects on human rights. Only 10 of the 24 reports that mentioned human rights addressed it adequately (score 3 or 4). The average score for this quality criterion is among the lowest and only one report described the issue well enough to receive a score of 4.

The cross-cutting issue of gender equality and women’s rights is the only cross-cutting issue on which a

majority of the reports that did address the issue were of adequate or good quality. However, this majority is small (53%) and only six reports were assessed to fully integrate gender equality and women’s rights issues without any gaps. Additionally, very few reports presented gender-segregated data and few reports included information about the gender of interviewees. Gender considerations were rarely included in the presentation of methodology and analysis. At least one decentralised evaluation ignored an explicit request for attention to gender considerations in the terms of reference.

The average score for integration of environment and climate change issues was 2.14. In one-third of the evaluations, the terms of references did not ask for integration of this cross-cutting issue; 11 reports ignored the request in the terms of references to integrate it. Half of the reports that considered climate and environmental issues were assessed as doing this with adequate or good quality. Only three reports were assessed to have fully integrated climate and environmental issues. In several cases, the raters noted that the intervention had a strong link to climate or environmental issues but that these issues were not integrated in data collection, findings, conclusions and recommendations to the extent that might be expected.

‘Integration of anti-corruption issues’ – for example, by discussing corruption risks or measures taken to avoid corruption – had the second lowest average score. In all, 34 reports, or 62% of the decentralised evaluations assessed, did not consider anti-corruption issues. The terms of references of 13 of these reports did not ask for integration of anti-corruption issues; in 21 of these evaluations, anti-corruption issues should have been integrated but were not.⁴⁷ The remaining 21 reports mentioned anti-corruption, but only 7 reports (13% of the total) integrated the issue well enough to be rated as adequate or good quality (scores 3 or 4). Interesting examples include one report that mentioned corruption as an issue in the description of context but did not discuss the issue further. Another report contained an observation indicating a risk that salaries may have been double paid but did not remark further on this.

⁴⁷ These 21 reports include reports where the terms of reference explicitly stated that the issues should be addressed as well as reports where the terms of reference did not specifically mention the issues but where the raters deemed the issues to be relevant.



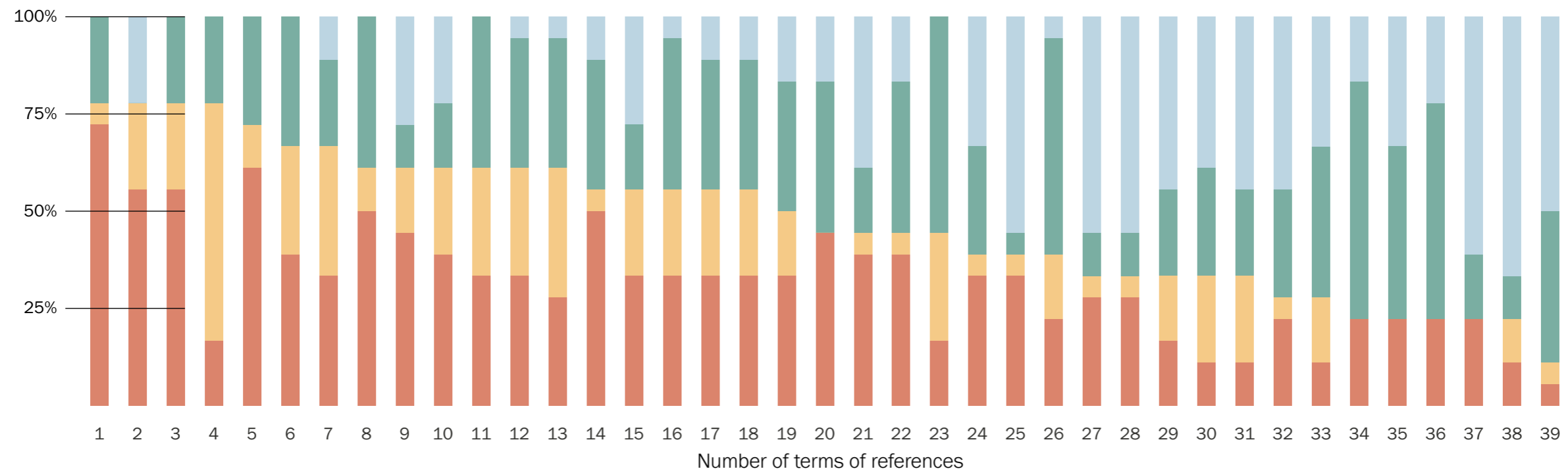
3.4 Terms of Reference Quality Criteria

Terms of references were rated using the same four-point grading scale as the decentralised evaluation reports, but without the detailed scoring guidance that was provided for rating the reports. The team of

raters instead used the general scoring guidance and descriptions provided in the Guidance Manual (see Table 1 in Chapter 2). It should be noted that for terms of references, the option of ‘not applicable’ is not available. As shown in Figure 10, the distribution of scores on quality for terms of references is similar to

that for reports (Figure 2).⁴⁸ Just over half of the terms of references (22 of 39) were of good or adequate quality (scores 3 or 4) on at least half of the quality criteria and only 6 terms of references were rated as of good or adequate quality on at least three-fourths of the quality criteria. As for reports, all terms of references had at least one quality criterion with score of 1. Eight terms of references did not have any criterion with score of 4.

Figure 10: Distribution of Scores: Terms of References



Score 1 (red), Score 2 (yellow), Score 3 (green), Score 4 (blue)

Note: Each of the columns represents the scores of one terms of reference. The figure is sorted by the total percentage of scores 1 and 2 for each of the 39 terms of references. The colours of the columns correspond to the quality scoring scale presented in Table 1.

48 Figure 15 in Annex 3 presents the distribution of scores across terms of references without sorting, and with identification numbers linking them to their respective report.



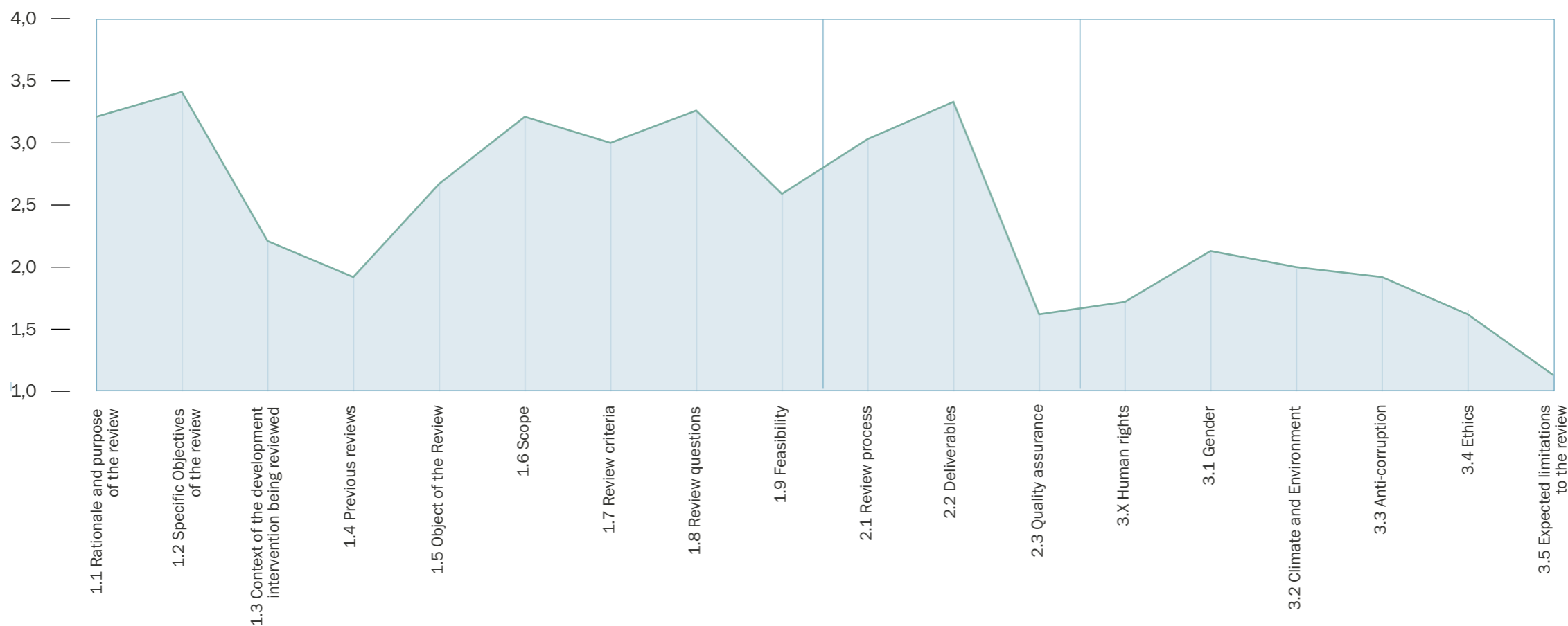
Figure 11 shows the average score for each of the quality criteria for terms of references (Figure 3 is the comparative illustration of scores for reports) and Figure 12 shows the distribution of scores for each quality criterion. (For reports, these are illustrated in

separate figures for each quality area). It should be noted that there are only three quality areas for terms of references (versus five for reports) and that only some quality criteria are the same for reports and terms of references. Figure 11 shows a similar pattern to that

for reports, with the highest average scores being just over 3 and with low average scores for context, previous reviews, cross-cutting issues, ethics and limitations.

Figure 11: Average Scores: Terms of Reference Quality Criteria

Note: Each segment separated by vertical blue lines covers one of three different quality areas. The green curve marks the average score for each quality criterion within the quality areas.



Quality criteria that describe the assignment (quality area 1, the first segment in Figure 11) have relatively high quality with several average scores above 3. The two criteria, ‘context of the development intervention’ and ‘previous reviews’, have lower average scores, but these still represent an improvement over the findings of Evaluation Department (2017). Only 41% of the terms of references describe the context well enough to be rated adequate or good (scores 3 or 4) and 56% of the terms of references give an adequate or good (scores 3 or 4) description of the object of the evaluation. A majority of the terms of references (62%) did not mention previous reviews or evaluations. The criterion ‘feasibility’ refers to whether the decentralised evaluation is feasible given the criteria, questions and resources made available. Several terms of references did not specify resources available, time line, report length, etc., making it difficult to assess feasibility and rate this criterion.

In quality area 2, the quality criterion ‘review process’ includes description of phases, deadlines and deliverables as well as distribution of roles and responsibilities. It was sufficiently well-described to merit a score 3 or 4 in 72% of the terms of references.

Notably, several terms of references did not mention an inception phase. Most terms of references focused on the role and responsibilities of the evaluator, but did not mention the commissioner’s responsibilities. Deliverables were adequately or well-described in 33 terms of references (85%). However, a large number of reports exceeded the requested number of pages for reports and executive summaries; this suggests that instructions regarding deliverables were often ignored by both evaluators and commissioners. Less than a third of the terms of references described required quality assurance procedures and only 9 terms of references, compared to none in Evaluation Department (2017), provided an adequate or good (scores 3 or 4) descriptions. These findings may partly explain the many weaknesses identified in the decentralised evaluation reports.

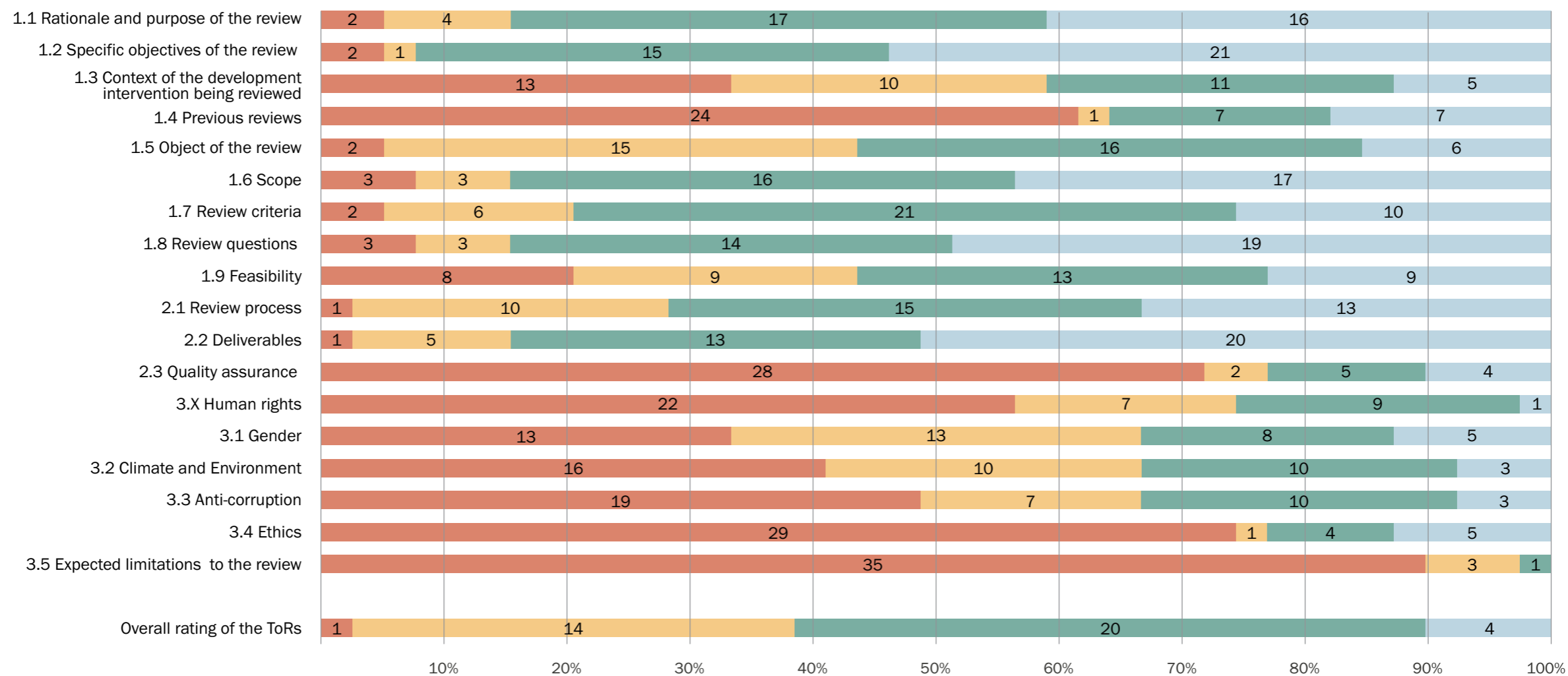
Quality area 3 includes cross-cutting issues, ethics and limitations. Cross-cutting issues were often summarily treated. Many terms of references included a separate evaluation question requesting the evaluation to “assess impact on cross-cutting issues”. Other terms of references selected one or more cross-cutting issues that the evaluation should assess. Very few terms of

references were more specific or asked for a more holistic approach. For example, one terms of reference required gender to be considered in the composition of the evaluation team. Another requested a gender lens to be applied. As in the evaluation reports, ethical issues and expected limitations were rarely even mentioned and, together with quality assurance, have the lowest average scores. Of the 39 terms of references assessed, 29 did not mention ethical issues and 35 did not mention expected limitations. Nonetheless, the scores on ethics and limitations are improvements over the results in Evaluation Department (2017).

The ‘overall rating of the terms of references’ (last row in Figure 12) shows the distribution of the raters’ overall assessments of the 39 terms of references: 1 was rated to be of overall poor quality, 14 were rated to be of less than adequate quality, 20 were assessed to be of adequate quality and only 4 were rated as of good quality. The average score for ‘overall rating of the terms of references’ was 2.75, which is slightly higher than the average of all quality criteria scores (2.44).



Figure 12: Distribution of Scores: Terms of Reference Quality Criteria



Score 1 (red), Score 2 (yellow), Score 3 (green), Score 4 (blue)

Note: The colours of the bars correspond to the quality scoring scale presented in Table 1 and the numbers are the number of terms of references that received the respective score. The score n/a is not available for terms of references.

Discussion and Conclusions

Nearly half of the reports and the terms of references had poor or less than adequate quality on at least half of the quality criteria.

A key conclusion is that the overall quality of both terms of references and decentralised evaluation reports is low. All terms of references and all but one of the 55 decentralised evaluation reports assessed for this report were assessed to be of poor quality on at least one criterion, and nearly half of the reports and the terms of references had poor or less than adequate quality on at least half of the quality criteria. None of the reports or terms of references were of adequate or good quality on all quality criteria. Only 16% of reports and 15% of terms of references had adequate or good quality on at least 75% of the quality criteria.

The raters found that information describing the assignment was of relatively good quality in both terms of references and reports, partly reflecting the fact that this information was often more or less a copy of the language in the respective terms of reference. In reports, information about rationale and users was often missing and some confusion regarding the term 'scope' is evident. The description of context often missed several important aspects, for example

cross-cutting issues. Previous reviews were often not described, indicating that prior knowledge was not fully used.

Methodology is the weakest quality area. The evaluation reports rarely described an analytical framework. The description of methods often consisted of a list of data collection methods rather than a description and justification of methods for both data collection and analysis. Primary sources were inconsistently referred to, if at all, and sampling and selection strategies were rarely presented. Many reports showed a lack of critical assessment of data, evidenced by the lack of both triangulation and discussion of robustness and limitations of data. Furthermore, ethical issues were mentioned in very few terms of references and evaluation reports.

The application of evaluation criteria is uneven. The reports showed higher quality on understanding and application of relevance, effectiveness and sustainability than on coherence, efficiency and



impact. Many reports revealed a poor understanding of efficiency, with the result that recommendations were often made without taking efficiency into consideration.

The raters found that in a majority of reports, the evaluation questions were answered. In some reports, the evaluators opted to include all OECD DAC evaluation criteria instead of focusing on the ones selected by the evaluation commissioners. Programme logics, theories of change or results chains were not fully assessed and rarely used to develop approach and methodology. In some reports, the distinction between evidence and findings or between findings and conclusions was not clear. In nearly half of the reports, the line of evidence was not fully clear.

Cross-cutting issues were often summarily treated in both terms of references and evaluation reports, and they were very rarely considered in the description of context or in evaluation design. Anti-corruption was especially poorly covered.

Based on the findings and this summary of weaknesses, the following problem areas are identified.

Development, Description and Justification of Approach and Methodology

Overall, issues relating to methodology were of relatively poor quality. These range from identifying an appropriate analytical framework and selecting methods for data collection and analysis to selecting sources of information and presenting the tools used to collect data. There are shortcomings in presentation, justification, and linking to evaluation criteria and questions. These shortcomings not only increase the risk that the evaluation methodology is not suited to the context, intervention and evaluation questions it is supposed to answer. They also limit transparency and make it difficult for the reader to correctly interpret and assess the merit of findings and conclusions.

Critical Reflection and Transparency Regarding Sources of Information

Several of the identified weaknesses point to a lack of critical reflection and transparency regarding how evidence is selected, presented and analysed. Examples include poor use or referencing of primary sources, lack of information about how sources were selected, and lack of transparent comparison of evidence from different sources (triangulation). Limitations, if mentioned, were mainly listed without comment as to how they may affect the robustness

of results. Ethical issues were not mentioned in most reports, either with regard to stakeholder integrity or in relation to sources of bias and reliability of evidence. The terms reliability, validity and robustness were rarely mentioned. These omissions limit the reader's opportunity to assess the robustness of evidence – if what is presented as evidence can be assumed to present a correct picture of reality – and hence if conclusions and recommendations can be trusted.

Efficiency, Anti-corruption and Financial Management

Both efficiency and anti-corruption affect how economically resources (funds, expertise, time, etc.) are converted into results.⁴⁹ The reports contain several examples of failed attempts to assess efficiency and few attempts to integrate anti-corruption. Several terms of references did not include anti-corruption as a cross-cutting issue even though it was relevant given the context or type of intervention. This lack of attention to efficiency, corruption risks and anti-corruption measures implies that corruption and inefficient use of resources may continue undetected and that recommendations involving large sums of money are made without proper analysis of these aspects.

⁴⁹ OECD 2019, p. 10.



Conceptual Confusion

Several of the concepts used in evaluations were not fully understood or were incorrectly used. As some concepts, such as scope, have multiple interpretations, communication about them may be unclear. Other concepts – for example, evidence, findings and conclusions – are well-understood but were not clearly separated in all reports. Finally, there are concepts that decentralised evaluations need to include in both evaluation processes and reports but that were missing or not well applied in many of the reports assessed. These include evaluation approach, triangulation, reliability, validity and robustness of data, and, last but not least, efficiency.

Use of Existing Information

Few terms of references and decentralised evaluation reports mention previous reviews and evaluations, and even fewer reports use these or follow up on

their findings or recommendations. Monitoring and evaluation data are more frequently used, although often without assessing the quality of the data or systems used to collect them. References to research literature were rare except in the context or background sections, where in some cases, they were included at the expense of more relevant information. This implies that a substantial amount of potentially useful and important information is not taken into consideration. The opportunity to learn from and build on previous evaluations and research is lost, and resources may be spent on collecting the same information a second time.

Adherence to, Use of and Reference to Terms of Reference

The raters found multiple examples of reports that ignored instructions in the terms of reference regarding the length of the report and executive summary,

integration of cross-cutting issues, and presentation of data and sources. A number of reports shifted the focus of the evaluation by adding evaluation criteria or ignoring evaluation questions. Such changes affect both quality and usefulness of the evaluation and imply that resources may be spent on pursuing the wrong issue. There is potential for evaluation commissioners to be stricter in demanding adherence to terms of references and thereby to contribute to improving the quality of decentralised evaluations.



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).
- Evaluation Department (2017). The Quality of Reviews and Decentralised Evaluations in Norwegian Development Cooperation. Evaluation Department report 1/2017, Norad.
- OECD (2019). OECD Development Co-operation Peer Reviews: Norway 2019, OECD Development Co-operation Peer Reviews, 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.
- OECD DAC (2019), Better Criteria for Better Evaluation Revised Evaluation Criteria Definitions and Principles for Use, OECD DAC Network on Development Evaluation. OECD Publishing, Paris.
- OECD DAC (2016). Evaluation Systems in Development Cooperation: 2016 Review. OECD Publishing, Paris.
- Overseas Development Institute (2006). Evaluating Humanitarian Action Using the OECD-DAC Criteria: an ALNAP Guide for Humanitarian Agencies. Overseas Development Institute, London.



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 evaluations⁵⁰ of development projects and programmes are an important source of information about the results of Norwegian development cooperation.⁵¹ Credibility and utility of these reviews and decentralised evaluations is therefore important.

⁵⁰ Hereafter mainly referred to as 'reviews and decentralised evaluations'.

⁵¹ 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.

Achieving adequate quality of decentralised evaluations is a challenge in many agencies.⁵² Therefore, many agencies, both bilateral donors and multilateral organisations, have institutionalised an external quality assessment mechanism to improve quality.⁵³ 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 and decentralized evaluations. The assignment will be a first step to establish this.

⁵² OECD DAC (2016) Evaluation Systems in Development Cooperation: 2016 Review. OECD Publishing, Paris.

⁵³ 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.

Reviews in the Norwegian Aid Administration

The quality of reviews and decentralised evaluations commissioned by the Norwegian aid administration⁵⁴ 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 1/2017.⁵⁵

The Evaluation department report 1/2017⁵⁶ 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

⁵⁴ 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.

⁵⁵ 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.

⁵⁶ Evaluation department Norad report 1/2017 'The Quality of Reviews and Decentralised Evaluations in Norwegian Development Cooperation'.



evaluations and decentralized evaluations is not made available to others.⁵⁷

Guidance for why, when and how to undertake reviews and decentralised evaluations is given in the GMM and requirements are specified in the rules⁵⁸ for each grant scheme. A review, as defined in the Grant Management Manual (GMM)⁵⁹ 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, Norad⁶⁰), implementing partners/grant recipients, and other agencies/co-sponsors. An estimated 60–70 reviews and decentralised evaluations are undertaken per year.⁶¹ All reviews and decentralised evaluations and evaluation reports shall be submitted to the evaluation portal⁶², 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.⁶³ Accompanying guidance material emphasise systematic use of evaluations as a source of management information and learning.⁶⁴

2. PURPOSE AND OBJECTIVES

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

57 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.

58 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.

59 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.)

60 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.

61 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.

62 <https://evalueringsportalen.no/>

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

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



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 published⁶⁵ 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, Norad⁶⁶ 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

⁶⁵ Published on the Evaluation Portal within 31st December each year.

⁶⁶ 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.

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 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 standards⁶⁷, developed for the previous evaluation of the quality of reviews and decentralised evaluations commissioned by the Evaluation

⁶⁷ OECD Development Assistance Committee <http://www.oecd.org/dac/evaluation/qualitystandardsfordevelopmentevaluation.htm>



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.⁶⁸ 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)).

⁶⁸ 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.

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.

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: Data Collection Tools

SCORING TEMPLATE FOR EVALUATION REPORTS

General information		Report Id (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, <i>approx.</i> hours
		Type of report (mid, end, etc.)
		Terms of reference were also reviewed (Y/N)

Key quality criteria	Quality statements	Score	Justification comments
Quality area 1: summary, style and structure			
1.1 Executive summary	The review ⁶⁹ contains an executive summary. It is complete and concise. It provides an accurate summary of the report, highlighting the rationale, purpose and specific objectives of the review/evaluation, the methodology used and the main findings, conclusions, and recommendations.		
1.2 Style and structure	The structure of the report allows for a clear flow of information from beginning to end. Each section builds on the previous sections with no jumps or gaps in information. The report is clearly written and properly edited.		

⁶⁹ Please note that the Guidance Manual uses the term 'review' to represent both reviews and decentralised evaluations.

Key quality criteria	Quality statements	Score	Justification comments
Quality area 2: Review purpose, objectives, and scope			
2.1 Rationale and purpose of the review	<p>The rationale, purpose, intended users and intended use of the review are stated clearly, addressing:</p> <ul style="list-style-type: none"> Why is the review or evaluation being undertaken? Why at this particular point in time? For whom is it undertaken? How is it to be used (i.e., for learning and/or accountability functions)? 		
2.2 Specific objectives of the review	<p>The specific objectives of the review clarify what the review aims to find out. Any modification to the specific objectives stated in the terms of reference is explained</p>		
2.3 Context of the development intervention	<p>The review describes relevant contextual information to the development intervention: Policies, objectives and strategies of the implementers. The development context, including socio-economic, political, cultural factors that are significant to the object of the evaluation. Key issues pertaining to Norway's cross-cutting themes (human rights; women's rights and gender equality; climate and environment; and anti-corruption) where applicable.</p>		
2.4 Review object	<p>The description of the intervention includes: The time period, budget, geographical area. Components of the intervention. Expected outcomes. Stakeholders. Organizational set-up/implementation arrangements. A summary of the intervention logic. Discrepancies between the planned and the actual implementation of the development intervention are explained.</p>		
2.5 Scope	<p>If the review scope encompasses the entire intervention, this is stated in the report. If the scope is limited to a subset of the intervention, that subset is described in addition to the intervention other dimensions to be covered by the review are also identified, if applicable. Modifications to the review scope established in the terms of reference are explained.</p>		



Key quality criteria	Quality statements	Score	Justification comments
2.6 Review criteria and questions	<p>The review should apply the agreed DAC criteria for evaluating development assistance (relevance, coherence, efficiency, effectiveness, impact and, sustainability and coherence) and Norway's cross-cutting themes (human rights, gender equality, climate and environment, and anti-corruption) unless alternative criteria and questions are clearly defined in the terms of reference.</p> <p>Questions are clear, specific, and answerable.</p> <p>Any modifications from the criteria and questions presented in the terms of references are explained and justified.</p>		
2.7 Previous reviews and decentralised evaluations	<p>Key findings and recommendations stemming from relevant previous reviews and decentralised evaluations / evaluations are mentioned.</p> <p><i>(Comment from The Evaluation Department in Norad: Keyword here is “mentioned” – if they state there are no previous reviews and decentralised evaluations, they have mentioned it. If they don't even say there are no previous reviews and decentralised evaluations, they have not mentioned it and the score should be 1.)</i></p>		
Quality area 3: Methodology			
3.1 Description of the design	<p>The report describes: The review/evaluation approach (conceptual framework). The review/evaluation design.</p>		
3.2 Sources of evidence	<p>Secondary data is referenced and primary data sources are clear: The sources of information used (documents, respondents, administrative data, literature, etc.) are clearly referenced and is referencing is consistent. The sampling and selection strategies in relation to specific data collection tools and approaches are clearly described.</p>		
3.3 Description of methods	<p>The review report describes (in report or annex): Instruments/techniques used for data collection, including those used to collect gender-sensitive data and information. How tools/techniques for data collection where used, or applied. Data analysis methods, including analysis of gender-sensitive data and information.</p>		
3.4 Monitoring and evaluation	<p>The strengths and weaknesses of monitoring and evaluation data/systems are described. The review makes use of the existing monitoring and evaluation data.</p>		



Key quality criteria	Quality statements	Score	Justification comments
3.5 Methodological appropriateness	The review methodology (including approach, design, methods for data collection, analysis and sampling) is appropriate given the review purpose, objectives and approach and is well justified. Methods are linked to and appropriate for each review question.		
3.6 Methodological robustness	Evidence is triangulated and the reliability of data is assessed.		
3.7 Limitations and challenges	The review report describes any limitations in process, data sources and sampling/samples, data collection and data analysis as well as their implications in terms of validity and reliability. Limitations regarding the representativeness of the sample for interpreting review results are described. Any obstruction of a free and open review process which may have influenced the findings is described.		
3.8 Ethics	Ethical issues such as privacy, anonymity, do-no-harm, inclusion/exclusion, and cultural appropriateness are described and the approach taken by the review to addressing them is described. Ethical safeguards are described and are appropriate for the issues identified (e.g. protection of confidentiality; protection of rights; protection of dignity and welfare of people; Informed consent; Feedback to participants)		
Quality area 4: Application of OECD DAC evaluation criteria			
4.1 Relevance	The report correctly interprets and assesses relevance in the context of the initiative. It refers to the extent to which the intervention is suited to the priorities and policies of the target group/recipient and donor.		
4.X Coherence	The report correctly interprets and assesses coherence. It refers to how well the intervention fits, the compatibility of the intervention with other interventions in a country, sector or institution.		
4.2 Effectiveness	The report correctly interprets and assesses effectiveness. It assesses the extent to which the intervention has met or likely to meet its objectives, and whether it is managing risk well.		
4.3 Efficiency	The report correctly interprets and assesses efficiency. It judges if the least costly resources possible are used in order to achieve the desired outputs. It may consider also whether alternative approaches would have produced the same results for less resources.		



Key quality criteria	Quality statements	Score	Justification comments
4.4 Sustainability	The report correctly interprets and assesses sustainability. It assesses the extent to which the benefits of the intervention are likely to continue after donor funding has been withdrawn. Projects need to be environmentally as well as financially sustainable.		
4.5 Impact	The report correctly interprets and assesses impact. It assesses the extent to which the initiative is likely to or has begun to attain its longer-term goals beyond the life of the intervention		
Quality area 5: Analysis, data, findings, conclusions, lessons learned and recommendations			
5.1 Review questions answered	The report answers all the questions detailed in the terms of reference for the review. The questions from the terms of reference, as well as any revisions are documented to enable readers to assess whether the review team has sufficiently addressed the questions and met the review objectives.		
5.2 Programme logic	The theory of change/programme logic is assessed in a comprehensive manner, and any gaps are identified. The theory of change/programme logic is assessed against relevant literature/evidence. A description of the assumptions underlying the theory of change/programme logic is included.		
5.3 Findings	Findings flow logically from the analysis of data, showing a clear line of evidence. Triangulation has been used to underpin findings. Gaps and limitations in the data are explained and the likely impact on the analysis assessed.		
5.4 Causal Inference	Findings clearly distinguish outputs, outcomes and impacts (where appropriate) and demonstrate the progression from implementation to results. Attribution and/or the extent of contribution of the intervention to expected outcomes is discussed. There is an exploration of other factors outside the intervention which may have influenced or caused achieved outcomes.		
5.5 Conclusions	Conclusions present reasonable judgments based on findings and substantiated by evidence and analysis. They add value to the findings, identifying priority issues, pertinent to the object and purpose of the review.		
5.6 Recommendations	The report contains clear, relevant, targeted and actionable (timed and prioritized) recommendations. Recommendations are grounded in the evidence and follow logically from the conclusions.		



Key quality criteria	Quality statements	Score	Justification comments
5.7 Lessons learned	If present, lessons follow logically from the conclusions. Lessons should only be drawn if they represent contributions to general knowledge.		
5.8a Integration of human rights	Human rights issues (if requested in the terms of reference) inform the findings, conclusions, recommendations and lessons as appropriate.		
5.8b Integration of gender equality and women's rights issues	Gender equality and women's rights issues (if requested in the terms of reference) inform the findings, conclusions, recommendations and lessons as appropriate.		
5.8c Integration of climate and environment	Climate and environment issues (if requested in the terms of reference) are integrated where appropriate into the findings, conclusions, recommendations and lessons.		
5.8d Integration of anti-corruption	Anti-corruption issues (if requested in the terms of reference) are integrated where appropriate into the findings, conclusions, recommendations and lessons.		
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.		
Recommendations	Recommendations made in the review, that go beyond programme level.		
Room for the scorer's comments and reflections, if any			
General reflections on the review	General reflections on the evaluation/review, key things missing from the report, good practise identified, positive outliers, etc.		
Comments about the scoring process			
Scoring process	Reflections on the tools used in the assessment (the scoring templates), useful tips, comments, questions etc.		



SCORING TEMPLATE, TERMS OF REFERENCES

General information	Report Id (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	Rating	Justification
1. Review purpose, objectives, object and scope			
1.1 Rationale and purpose of the review	The rationale, purpose, intended users and intended use of the review are stated clearly, addressing: <ul style="list-style-type: none"> – Why is the review being undertaken? – Why at this particular point in time? – For whom is it undertaken? There is specificity about the intended audience (beyond simply identifying institutions) – How is it to be used (i.e. for learning and/or accountability functions)? 		
1.2 Specific objectives of the review	The specific objectives of the review clarify what the review aims to find out		
1.3 Context of the development intervention being reviewed	The terms of references contain a brief description of the context of the intervention being evaluated. This may include: <ul style="list-style-type: none"> – policy context (Norway's and partners' policies, objectives and strategies) – development context, including socio-economic, environmental, political, cultural factors – key issues pertaining to Norway's cross-cutting themes (i.e. women's rights and gender equality; climate and environment; and anti-corruption). 		
1.4 Previous reviews and decentralised evaluations	The terms of reference states whether previous reviews and decentralised evaluations exist, and if applicable, identifies relevant issues		



Key quality areas	Quality statement	Rating	Justification
1.5 Object of the review	The development intervention being reviewed (the review object) is clearly described, including: <ul style="list-style-type: none"> – period – budget – geographical area – Intervention logic/theory of change/logic model – expected outcomes – stakeholders – organizational set-up 		
1.6 Scope	The terms of references clearly define what will and will not be covered by the review, including: <ul style="list-style-type: none"> – What aspect/dimensions of the intervention. – the time period – the geographic coverage 		
1.7 Review criteria	Based on the review mandate, the terms of reference identifies the relevant criteria (OECD DAC, cross-cutting themes and issues) for the review: <ul style="list-style-type: none"> – OECD DAC: relevance, coherence, efficiency, effectiveness, impact and sustainability – Cross-cutting themes: human rights, women’s rights and gender equality; climate and environment; and anti-corruption 		
1.8 Review questions	The questions are customized and rendered specific to users’ (as defined in the rationale and purpose section) information needs.		
1.9 Feasibility	The scope of work proposed by the terms of reference is feasible given the timeframe and resources provided. The terms of reference contain a limited/ prioritized number of review questions that are clear and relevant to the object and purpose of the review.		



Key quality areas	Quality statement	Rating	Justification
2. Review process and quality assurance			
2.1 Review process	The review terms of reference clearly explains what is expected of the Consultant in terms of: <ul style="list-style-type: none"> – having an inception stage – data collection and validation – preparing the review or review report – Roles and responsibilities of the team members (consultants) and of Norad/ Norwegian Ministry of Foreign Affairs/ Embassy/ Partner (review manager) are defined and appropriate to the review objectives 		
2.2 Deliverables	The review terms of reference identifies the mandatory deliverables and milestones: <ul style="list-style-type: none"> – inception report (if applicable) – debriefing / validation sessions – draft and final review report – presentation of the report (optional) The schedule identifies the key phases of the review.		
2.3 Quality assurance	The terms of reference specify that the review will follow professional norms and standards, including OECD DAC. Provisions for quality assurance mechanisms are included in the terms of reference.		
3. Overarching and cross-cutting criteria			
3.X Human rights	Human rights are reflected in the terms of reference where appropriate (context, design, questions around effectiveness and impact)		
3.1 Gender	Gender dimensions and women's rights are explicitly addressed in all relevant parts of the terms of reference (context, questions, approach, design, methods, team composition)		
3.2 Climate and environment	Climate and environment dimensions are reflected in the terms of reference where appropriate (context, design, questions around effectiveness and impact)		
3.3 Anti-corruption	Anti-corruption issues are reflected in the terms of reference (e.g. as part of risks or context)		



Key quality areas	Quality statement	Rating	Justification
3.4 Ethics	Ethical considerations (consent, protection, participation, independence) and requirements are explicitly addressed		
3.5 Expected limitations to the review	Expected limitations to the review are identified (methods, sources of info, disaggregated data, time, budget)		
Overall rating			
Overall rating of the terms of reference	The terms of reference provide a sound basis for the review, that will guide the review manager and team on how to fulfil effectively the objectives of the review		
Good practise	List any examples of good practice		
General comments	Assessor's general comments		



Annex 3: Presentation of Data

Table 3: Average Scores, Report Quality Criteria

	Average	Standard deviation
Quality area 1: Summary, style and structure		
1.1 Executive summary	2.71	0.77
1.2 Style and structure	2.87	0.79
	Average	2.79
Quality area 2: Evaluation purpose, objectives, and scope		
2.1 Rationale and purpose of the evaluation	2.80	0.88
2.2 Specific objectives of the evaluation	2.96	0.89
2.3 Context of the development intervention	2.65	0.90
2.4 Evaluation object	3.02	0.77
2.5 Scope	3.02	0.77
2.6 Evaluation criteria and questions	2.82	0.81
2.7 Previous reviews and decentralised evaluations	1.94	1.12
	Average	2.75



	Average	Standard deviation
Quality area 3: Methodology		
3.1 Description of the design	2.18	1.01
3.2 Sources of evidence	2.45	0.80
3.3 Description of methods	2.25	0.96
3.4 Monitoring and evaluation	2.47	0.78
3.5 Methodological appropriateness	2.35	0.86
3.6 Methodological robustness	2.18	0.74
3.7 Limitations and challenges	2.07	0.87
3.8 Ethics	1.16	0.56
	Average	2.14
Quality area 4: Application of OECD DAC evaluation criteria		
4.1 Relevance	3.06	0.81
4.X Coherence	2.39	0.92
4.2 Effectiveness	3.29	0.73
4.3 Efficiency	2.46	0.91
4.4 Sustainability	2.96	0.80
4.5 Impact	2.44	0.91
	Average	2.83



	Average	Standard deviation
Quality area 5: Analysis, data, findings, conclusions, lessons learned and recommendations		
5.1 Evaluation questions answered	2.98	0.86
5.2 Programme logic	2.29	0.98
5.3 Findings	2.64	0.90
5.4 Causal Inference	2.73	0.84
5.5 Conclusions	2.84	0.78
5.6 Recommendations	2.84	0.83
5.7 Lessons learned	2.11	1.04
	Average	2.66
Cross-cutting Issues		
5.8a Integration of human rights	2.00	0.83
5.8b Integration of gender equality and women's rights issues	2.55	0.85
5.8c Integration of climate and environment	2.14	0.93
5.8d Integration of anti-corruption	1.71	0.85
	Average	2.12



Table 4: Average Scores, Terms of Reference Quality Criteria

	Average	Standard deviation
1. Evaluation purpose, objectives, object and scope		
1.1 Rationale and purpose of the review	3.21	0.82
1.2 Specific Objectives of the review	3.41	0.78
1.3 Context of the development intervention being reviewed	2.21	1.04
1.4 Previous reviews	1.92	1.23
1.5 Object of the Review	2.67	0.80
1.6 Scope	3.21	0.88
1.7 Review criteria	3.00	0.78
1.8 Review questions	3.26	0.90
1.9 Feasibility	2.59	1.06
	Average	2.83
2. Evaluation Process and Quality Assurance		
2.1 Review process	3.03	0.83
2.2 Deliverables	3.33	0.80
2.3 Quality assurance	1.62	1.05
	Average	2.66



	Average	Standard deviation
3. Overarching and cross-cutting criteria		
3.X Human rights	1.72	0.90
3.1 Gender	2.13	1.02
3.2 Climate and Environment	2.00	0.99
3.3 Anti-corruption	1.92	1.02
3.4 Ethics	1.62	1.10
3.5 Expected limitations to the review	1.13	0.40
	Average 1.75	
The rater's overall rating of the terms of reference	2.70	0.68



Figure 13: Distribution of Scores per Report

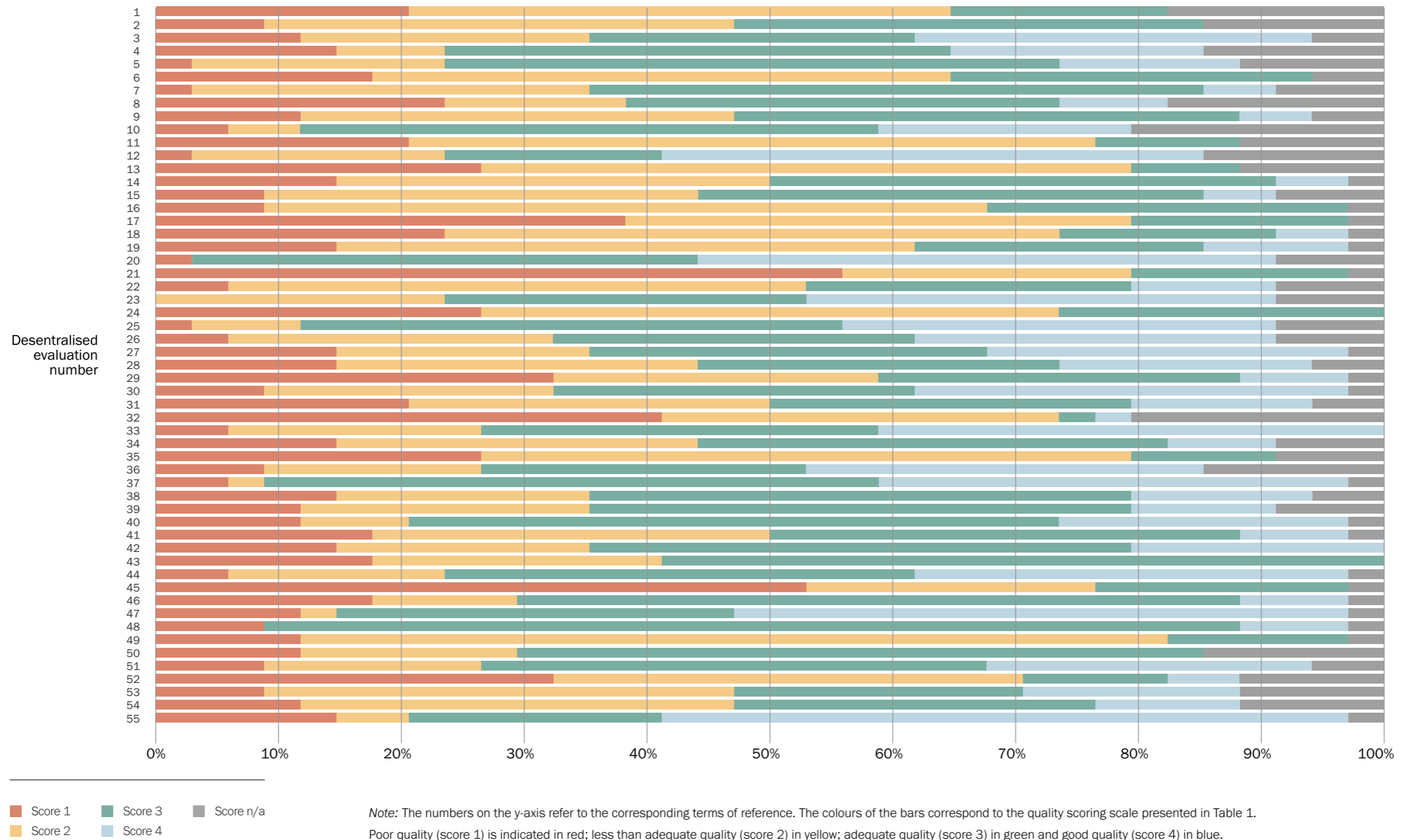
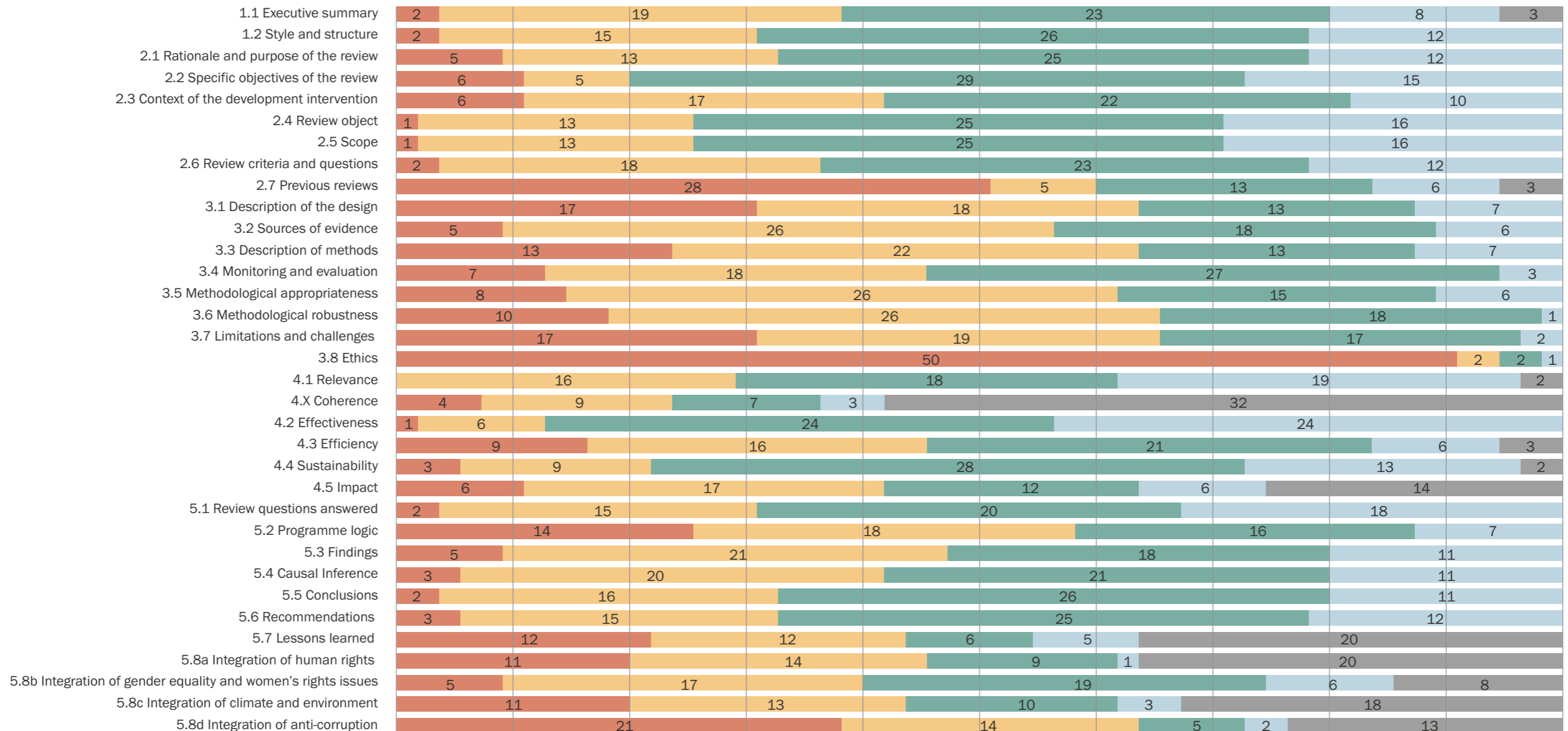


Figure 14: Distribution of Scores per Report Quality Criterion



Score 1 Score 3 Score n/a
 Score 2 Score 4

Note: The colours of the bars correspond to the quality scoring scale presented in Table 1 and the numbers are the number of reports that received the respective score.

Figure 15: Comparison of Average Scores for Reports Referred to as ‘Evaluations’ and ‘Reviews’

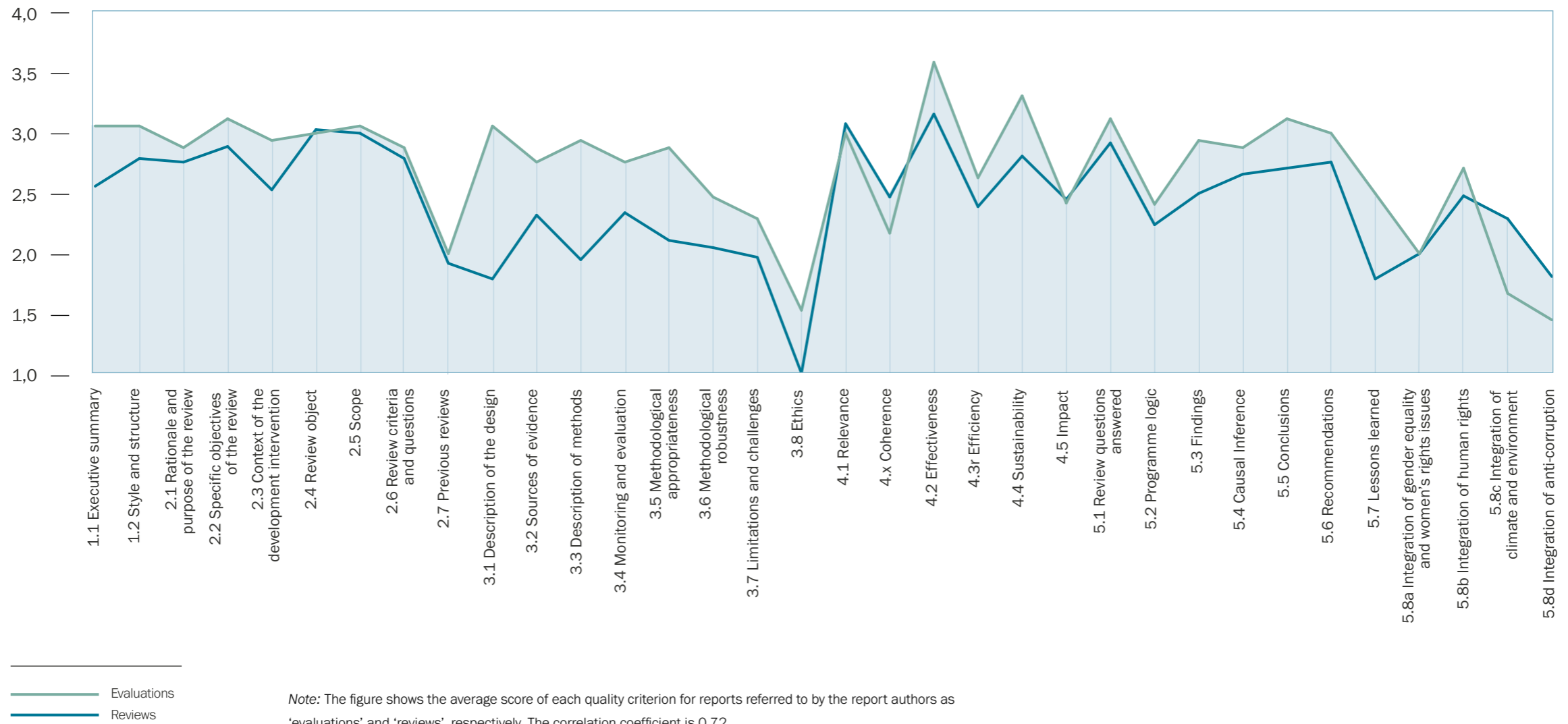
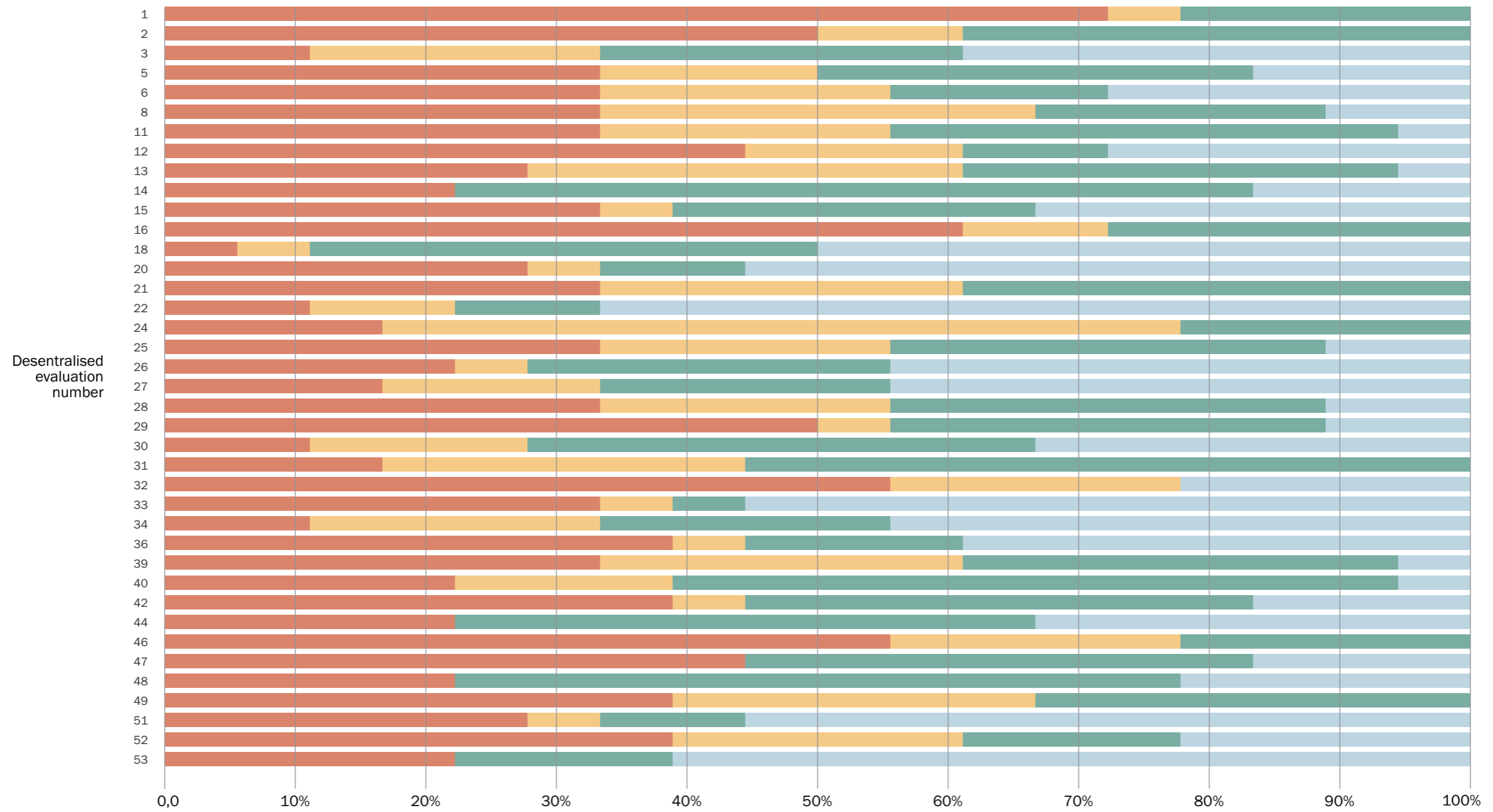


Figure 16: Distribution of Scores per Terms of Reference



Score 1 (red), Score 2 (orange), Score 3 (green), Score 4 (blue)

Note: The numbers on the y-axis refer to the corresponding evaluation report number. The colours of the bars correspond to the quality scoring scale presented in Table 1. The figure shows scores for terms of reference quality criteria. Note that only evaluations for which the raters had access to terms of references are included. A missing number on the vertical axis implies that this evaluation did not come with a terms of reference (for example, decentralised evaluation number 4, 7, 9, 10 etc.)



Annex 4: Best Practise Evaluations

This annex provides summaries of three of the 55 quality assessed reports with the highest average scores. The first report has the highest average score across all quality criteria. The second report has the highest average score on methodology criteria. The third has the second-highest score both overall and on methodology criteria. In addition to summarising the findings, Annex 4 give examples of what made these reports score high.

INDEPENDENT TERMINAL EVALUATION: REPUBLIC OF SUDAN – BUILDING INSTITUTIONAL CAPACITIES FOR THE SUSTAINABLE MANAGEMENT OF THE MARINE FISHERY IN THE RED SEA STATE

The report presents the main findings, conclusions and recommendations from an independent evaluation of a United Nations Industrial Development Organization (UNIDO) project in Sudan entitled, “Building institutional capacities for the sustainable management of the marine fishery in the Red Sea State” (UNIDO Project No.: 130130).⁷⁰

⁷⁰ UNIDO Independent Evaluation Division, 2018, Independent Terminal Evaluation: Republic of Sudan – Building institutional capacities for the sustainable management of the marine fishery in the Red Sea State. UNIDO, Vienna 2018. Evaluators: Andrew Young and Salih Suliman.

“The development goal of the project is to contribute to sustainable management of marine fisheries in the Red Sea State with the outcome that relevant institutions have strengthened their capabilities to develop and maintain a data base on fish stocks and fish landings. Outputs include four surveys (150 days in total) undertaken in the Red Sea of Sudan and the creation of a web-based centralized database of fisheries data.”⁷¹

UNIDO implemented the project and also managed the evaluation. The terms of reference are extensive and include a great deal of detail about the context, intervention and evaluation process. The objective of the evaluation is to assess the performance of the project against the OECD DAC evaluation criteria and provide short- and longer-term strategic recommendations to the project as a further phase is anticipated.

This is a well-written report with a logical structure and clear flow from evidence through to findings and conclusions and on to recommendations. It is well-structured with a good balance of background and context on one hand and assessment and analysis on

⁷¹ UNIDO Independent Evaluation Division (2018) p. viii.

the other. The report states clearly why the evaluation is undertaken, why it is undertaken at that point in time, and how and by whom it is to be used. The context is described well and contains only relevant information. However, contextual information relating to cross-cutting issues is missing.

It is explicitly stated that the entire project is to be included and that all activities and the entire results chain are to be assessed. The description of the intervention is brief but clear and includes details regarding budget, time period, etc. It also links the project to the Millennium Development Goals (MDGs) and Sustainable Development Goals:

“Environmental sustainability (MDG7) is at the core of the project but the project also has the longer-term potential to contribute to MDG1, the eradication of extreme poverty and hunger by improving food security and the opportunities for diversifying local economies and livelihoods in areas such as processing or export.”⁷²

The evaluation process is illustrated in a figure indicating feedback processes and presentation

⁷² Ibid. p. 10.



of preliminary findings to stakeholders. No specific conceptual framework is presented, but the report describes relatively well how the methodology was developed, including with references to a mid-term evaluation of the project and the UNIDO evaluation manual. It is also stated that the methodology was presented in an inception report. Evaluation questions and sources are linked and presented in an annexed evaluation matrix. A rating tool was applied to the overall project design and log frame; project performance including relevance, efficiency, effectiveness and sustainability; performance on cross-cutting issues; and performance of partners. The description of the rating tool is brief; additional detail on how this tool was used in analysing data would have been useful.

The terms of reference are very specific as to how the review is to be implemented. The evaluation management seems to have been actively involved in supporting the evaluation, and UNIDO offices and the National Project Coordinator assisted in the selection of interviewees. The latter is clearly stated, illustrating transparency, but there is no description of how interviewees were selected. Furthermore, the potential effect of UNIDO and the coordinator's involvement in the selection process is not mentioned, although

this is a potential source of bias that could have been discussed under limitations.

Monitoring and evaluation data are well used throughout the report, and the quality of this data is discussed. Documents are consistently referenced but references to interview data are somewhat vague as they refer to organisations rather than individuals. Information from different sources and methods is triangulated and used to inform findings. The following citation illustrates both triangulation and the vague referencing of interviews:

“Project progress reports indicated some significant milestones of the project and these were validated with UNIDO in Port Sudan, extensively with IMR [Institute of Marine Research of Norway] and with some counterparts.”⁷³

Ethical issues are not mentioned. The report comments on reliability in relation to the lack of quantitative data and the mitigation strategy. However, the report makes no overall comment on reliability and validity and presents only one limitation together with a mitigation strategy:

⁷³ Ibid. p. 17.

“As beneficiaries of the project were organizations and not large numbers of individuals no quantitative analysis was carried out. However the evaluation team considers that data collected from extensive qualitative questioning and cross referencing provided sufficient validation and triangulation by comparing multiple verbal responses with progress reports, project documents and a broad range of project literature.”⁷⁴

The application of OECD DAC evaluation criteria is appropriate and illustrates good understanding of the criteria. Relevance is assessed with respect to all levels and actors, from local to global, and includes discussions of relevance to policies and strategies at various levels, including donors. The project is assessed as highly relevant with no shortcomings. Effectiveness is correctly interpreted and adequately assessed. The report triangulates evidence from project reports with interview data and finds the following:

“The project is rated as Satisfactory regarding effectiveness with only minor shortcomings. The project has delivered a wide range of activities which clearly support anticipated outputs and the outputs should

⁷⁴ Ibid. p. iii.



lead to the outcome of strengthened capacities to develop and maintain data on fish stocks and fish landings in the RSS. The project has delivered outputs as expected, in a timely manner, and cost effectively.”⁷⁵

Efficiency is assessed together with coordination and management. The report discusses, for example, whether the distribution of work between organisations was appropriate:

“Efficiency is enhanced with distribution of tasks in line with comparative advantages: UNIDO managing logistics, transportation and procurement and IMR [Institute of Marine Research of Norway] delivering the implementation of training and technical transfer through work packages (WP) agreed and articulated under sub contract.”⁷⁶

The main challenge to both efficiency and effectiveness was the depreciation of the Norwegian kroner against the Euro, which resulted in the project receiving 11% less funding than anticipated. The evaluation describes how this was handled and finds that it did not affect

⁷⁵ Ibid. p. 6.

⁷⁶

achievement of the main outputs. Sustainability is well-covered for a range of components including financial and environmental aspects, ownership, and socio-political risks. Impact and challenges are well-covered. Unintended effects are noted and attribution is referred to. The report concludes the following:

“An impact on contributing to sustainable management of marine fisheries appears likely. With perhaps the most scientific and comprehensive stocktaking exercises being undertaken to date, the evident commitment of national partners and the ongoing interest from UNIDO and the IMR [Institute of Marine Research of Norway] it is assessed the possibilities to enhance impact are also evident. Capacities toward best practice data collection and analysis methodologies have [been] strengthened for three national counterpart organizations and impact is evident with greater working synergy between them. [...] Data outputs of the project have already been incorporated into new fisheries regulations and the 5-year strategy of the Industrial Modernization Programme of the Republic of the Sudan.”⁷⁷

⁷⁷ Ibid. p. ix.

The report follows the terms of reference and, with minor exceptions, answers the evaluation questions. The line of evidence flows logically with clear separation of findings and evidence, and the report makes a clear distinction between output, outcome and impact. The project’s contribution to intended results is demonstrated and other contributing factors are discussed (e.g. earlier projects, UNIDO’s experience, etc.). The report presents reasonable and relevant conclusions that are based on findings and add value and focus on priority issues. Recommendations are grounded in evidence, specific, realistic, targeted and actionable. The report demonstrates that the concept ‘Lessons learned’ is correctly interpreted. Lessons learned are well-formulated and of general interest, although some are more relevant for the organisation implementing the programme.

The report states that there are no negative human rights aspects of the project. Anti-corruption issues are not covered but also not asked for in the terms of reference. Climate and environment issues are relatively well-covered, including environmental risks and effects. Gender equality and women’s rights issues are included in several sections of the report, for example impact, unintended effects and gender sections. These inform findings but not conclusions and



recommendations. The following unintended effect was identified:

“While fishing is an exclusively male occupation in Sudan the project has enabled large numbers of women to become involved in fisheries management. Women are involved in overall management on the PSC [Project Steering Committee] and have a strong research presence in both the RSFRS [Red Sea Fisheries Research Station] and the URS-FMSF [University of the Red Sea State-Faculty of Marine Sciences and Fisheries]. In the Annual Survey observed by the evaluation, women also formed part of the survey team.”⁷⁸

ADVANCING AND SUSTAINING GENDER BASED GOVERNANCE IN MALAWI 2014 – 2018: END OF PROGRAMME EVALUATION

The report is the end evaluation of the programme entitled Advancing and Sustaining Gender Based Governance in Malawi (the GBG programme).⁷⁹ The UN Women Malawi Country Office implemented the

⁷⁸ Ibid. p. 25

⁷⁹ UN Women Malawi, 2019, End of Programme Evaluation: Advancing and Sustaining Gender Based Governance in Malawi 2014 – 2018, Malawi Country Office, Lilongwe. Evaluator: Hope Msosa.

GBG programme from 2015 to 2018; Royal Norwegian Embassy provided funding.⁸⁰ According to the evaluation report, the goal of the programme was to accomplish the following:

“position gender equality as central to all development processes in Malawi. The programme had three outcome areas, namely: gender equality dimension is mainstreamed in policies, strategies and budgets at all levels; enhanced capacity of Parliamentarian Women Caucus (PWC) and Standing Committees for gender sensitive oversight, representation and legislative function; non-state actors effectively influence gender agenda in Malawi. The total cost of the programme was estimated at United States Dollars (USD) 2, 076,089 and worked with selected institutions (state and non-state) in the National Gender Machinery (NGM).”⁸¹

This report is assessed as an example of good evaluation practice. It is accessible, complete with annexes and very well-structured. The executive summary is complete but, at eight pages, is also considered to be too long.

⁸⁰ Grant contract ATLAS 93275; Government of Norway reference Norway Grant MWI-14/0016.

⁸¹ UN Women Malawi (2019), p. v.

The rationale and purpose are briefly described in the evaluation. While the description of the specific objectives is limited but clear, there are some gaps in terms of how the report is going to be used. Evaluation criteria and questions (both OECD DAC evaluation criteria and criteria from the UN Women Global Evaluation Report Assessment and Analysis System) are discussed and described well in the main text, and an annexed evaluation matrix complements this discussion. There is a clear link to a mid-term evaluation, with findings referenced, presented and discussed in several places in the report. The background chapter provides a full and appropriate description of policy and socio-economic context and a clear and detailed description of the programme, including the intervention logic.

The five-page methodology chapter of the evaluation report includes separate sections on sample and sampling design, data analysis, and ethical, gender and human rights considerations as well as evaluation limitations. These provide the reader a good understanding of how the evaluation was implemented and justifies choices made. The approach of the evaluation is explained as follows:



“The evaluation adopted a gender responsive and human rights based approach in its design, tools and execution. It was grounded in key women’s rights frameworks, including CEDAW [Convention on the Elimination of all Forms of Discrimination Against Women] and the Beijing Platform for Action. It was also based on the principles of empowerment, participation of stakeholders, and inclusiveness. The evaluation emphasized the active participation of stakeholders. [...] Further, to adhere to the United Nations Evaluation Group (UNEG) evaluation quality standards, the evaluation used the Global Evaluation Report Assessment and Analysis System (GERAAS) for quality benchmarking.”⁸²

The methodology is assessed to be appropriate; methods are linked to evaluation questions in an annexed evaluation matrix. The report includes a description of the instruments for data collection and comments explaining why specific methods were used; a selection of tools is presented in an annex. Adding to the transparency, sources of information and sampling strategies are clearly described, as is the focus of questions in each category:

⁸² Ibid. p. 10.

“The sampling approach for the exercise was largely purposeful and at different levels. The first level of the sample included national level stakeholders who directly implemented and benefited from the programme. [...] The other level of the sample included district level stakeholders i.e. district councils which directly benefited from the programme. In that regard, five district councils were selected to participate in the exercise. [...] In the districts at least 5 council staff members were consulted. These staff members included those that directly received capacity building support through the programme. However, where necessary, other staff members i.e. those that did not participate in any capacity building exercise, were also consulted.”⁸³

The number of stakeholders consulted is stated, with institutions listed in the methods chapter and annex. However, individual respondents are not named. The report also describes how documents were sourced and used to inform the evaluation. Data analysis methods are clearly described, allowing the reader to assess credibility of findings:

⁸³ Ibid. p. 10.

“Data from all sources including desk review, field visits, and interviews was analysed as follows:

- a) The triangulation of data emerging from these divergent sources was used to ensure validity and reliability of the findings. Triangulating information was used to identify similarities and/or discrepancies in data obtained from different sources (desk review of documents and key informant interviews) and from different stakeholders (duty bearers, rights holders, etc.).*
- b) Content analysis was conducted on qualitative data collected through document review and key informant interviews. The programme indicators on the output level were used to measure the results and used to establish quantitative and qualitative changes over a period of time.*
- c) Comparative analysis of the results planned in the original programme document and subsequent programme reports i.e. [Midterm review] MTR, first year implementation report and final programme report.”⁸⁴*

⁸⁴ Ibid. p. 12.



The report describes limitations encountered during the evaluation process, mitigation approaches and remaining effects. It also describes ethical, gender and human rights considerations and the approaches taken to include these considerations in the evaluation design. However, it does not describe remaining shortcomings and how these may have affected the reliability of data.

The report comments on deficiencies in the intervention's monitoring and evaluation systems and then makes use of available data. Data sources are described and consistently referenced, although interviewees are not individually referred to. Information from different sources is presented to inform findings:

“The other key issue raised by most of the stakeholders consulted is the scope and period of the programme. Discussions with national level government stakeholders as well as council level stakeholders revealed that the programme needed to widen its scope in terms of support. For example, key local council officials especially in Salima and Dedza stated that gender responsive budget trainings needed to reach out to all council sectors in order to widely embed the knowledge and skills; and create a larger number of gender equality champions.”⁸⁵

Findings are presented with clear links to evidence and evaluation questions. The layout makes it very easy for the reader to follow the link from data to evidence and findings for each evaluation question. Evaluation questions are presented in boxes, followed by findings in bold text and a paragraph or two presenting data supporting the finding. Information from different sources is often compared. Figure 17 is an example.

Figure 17: [Example of the Layout in the Findings Chapter of UN Women Malawi \(2019\)](#)

Question 3: To what extent did stakeholders and beneficiaries participate in the design and implementation of the programme?

Finding 4: There was no indication that all stakeholders and beneficiaries specifically participated in the design of the programme; but rather largely in implementation.

Discussions with representatives/staff from NICE, MALGA, NSO and PWC revealed that their institutions were not directly consulted in the formulation of the programme. However, high level discussions were conducted with the MoGDSW (as the lead institution in the National Gender Machinery) in terms of the support required to push the gender equality agenda. These consultations and discussions took place as part of developing the broader UN Women country priority areas. Notwithstanding, the evaluation finds that most stakeholders participated in implementing the programme. As earlier indicated, UN Women mostly took a facilitative role, relying mostly on the partners themselves to push for activities. A good example is NSO which implemented a project to engender statistical reports.

Source: UN Women Malawi (2019), p. 18.

85 Ibid. p. 18.



The evaluation notes the GBG programme design was informed by an objective context analysis, but that a deeper causality analysis was lacking. The programme design also was found to be appropriate to meet certain needs of the stakeholders but, as illustrated by the finding shown in Figure 17, stakeholders and beneficiaries did not participate in the design of the programme.

The report contains a good discussion of the relevance of the intervention to the main stakeholders and beneficiaries, but its relevance to the Norwegian donor is not mentioned. The GBG programme was found to be strongly aligned to national structures and policy as well as to international and regional aspirations on gender equality. The programme ambitions were relevant and, to a large extent, the programme met the needs and priorities of the key implementing stakeholders.

There is a thorough presentation and analysis of effectiveness including risk factors. Each intended outcome is reviewed and achievement of milestones is assessed. Due to the lack of an updated logical framework, the assessment relies heavily on interview responses. The evaluation describes how questions were asked to capture this information. The evaluation found that the programme contributed to

the mainstreaming of a gender equality dimension in policies, strategies and budgets and that it successfully achieved the outcome related to building the capacity of the Malawi parliament for gender-sensitive oversight, representation and legislation. However, the third outcome – “non-state actors effectively influence gender agenda in Malawi” – was not widely achieved.

The report notes a lack of robust activity planning and monitoring systems, which made it difficult to assess cost effectiveness. It identifies mismatches between planned and actual activities on some outputs and untimely activity implementation. Substantial delays in disbursements of funds were noted, and these had a ripple effect on the respective projects under the programme. One implementing partner stated that activities were implemented at the “last minute” and at inopportune times due to these delays.⁸⁶ The report interprets the non-adherence to planning, monitoring and reporting as indicative of gaps in management, coordination and monitoring, but a more thorough analysis of such gaps is not made.

Sustainability of the results and benefits of the programme being evaluated is discussed and assessed

⁸⁶ Ibid. p. 32.

in a clear and credible way. According to the report, all stakeholders consulted during the evaluation stated that continuing the programme’s benefits after it is phased out will be challenging. The report also noted that while stakeholder engagement is likely to continue following its phase-out, the programme is unlikely to be scaled up, replicated or institutionalised.⁸⁷

Two cross-cutting issues are integrated in the evaluation. First, a human rights lens is applied, including discussion of duty bearers and duty holders. Second, gender equality and women’s rights are implicitly integrated as this issue is central to the object of the evaluation. The evaluation found that both gender equality and human rights considerations were integrated in the programme. However, such considerations were not always explicitly integrated in both the design or implementation, and where integrated, they largely reflected the duty-bearer perspective and not a right’s holder perspective.

All questions are answered in accordance with the terms of reference and an elaborate evaluation matrix. The programme logic is only partially assessed, as updates had not been made. The assessment of

⁸⁷ Ibid. p. 36.



findings is convincing and makes frequent reference to interviews, but it could have been improved by more triangulation and reference to documented evidence (if such evidence existed). The evaluation provides a good discussion of the programme's contribution towards expected outcomes but less of a discussion on impacts. Conclusions are reasonable and pertinent overall, and recommendations are targeted and justified but not especially concrete. Lessons learned follow logically from conclusions and recommendations and are generally applicable and clear, as illustrated by the following examples:

“Building a strong monitoring and evaluation framework, including a monitoring, evaluation and reporting plan, at the design of the programme would have helped in tracking progress; especially tracking higher level or outcome level results that could directly be attributed to the programme.”⁸⁸

“A clear programme exit strategy would have helped sustain the benefits and results of the programme. The strategy would have provided for immediate and future programme synergies with other organizations implementing gender equality initiatives. Most

⁸⁸ Ibid. p. 39.

stakeholders felt the programme was largely ‘one-off’ and could not link it with any further interventions supporting the national gender machinery.”⁸⁹

MID-TERM REVIEW OF THE RESULTS BASED PAYMENT TO THE CRGE FACILITY – PARTNERSHIP AGREEMENT FOR REDD+

The Norwegian project entitled Results Based Payment to the Climate Resilient Green Economy (CRGE) Facility⁹⁰ is intended to support the Strengthening for Forest Sector Development in Ethiopia Programme and assist Ethiopia in reaching the CRGE Strategy's goals for afforestation, reforestation and forest management and, consequently, increased carbon sequestration.⁹¹ Planned outcomes by 2020 include strengthened institutional capacity of the forest sector at all levels; promotion of forest conservation and development for their multiple benefits; facilitation of private sector involvement in forest development; promotion of science and innovation for enhancing sustainable forest

⁸⁹ Ibid. p. 39.

⁹⁰ Ibid. p. iii.

⁹¹ LTS International Limited, 2018, Mid-Term Review of the Result Based Payment to the CRGE Facility – Partnership Agreement for REDD+. Norad 2018. REDD+: United Nations Collaborative Programme on Reducing Emissions from Deforestation and forest Degradation.

management; and enhanced stakeholder engagement in forest development.

This mid-term review is an example of a well-structured report that clearly sets forth the data used and the potential and limitations of the data. The report seems to be useful in terms of providing both information about achievements so far (accountability) and information to be used for extension and scaling up (learning). It is obvious that the evaluation team came with good evaluation experience and subject knowledge. The evaluation also has well-designed terms of reference.

The report has a complete but long summary (of seven pages) that includes rationale and purpose, methodology, key findings and conclusions, and a good summary of recommendations. The report structure is close to ideal, with a good background chapter that includes a description of the programme and its basic log frame, a separate chapter for the purpose of the review, and a detailed methodology chapter. Findings are neatly arranged according to the OECD DAC evaluation criteria and the connected 31 evaluation questions, and these are followed by conclusions and recommendations that are clearly linked to findings. In addition, an array of annexes provide detailed data, information about sources, and data collection tools.



The rationale, purpose, scope, use and users are clearly described:

“The audience for the MTR consists in the first instance of the Norwegian Embassy in Addis Ababa and the Norwegian International Climate and Forest Initiative (NICFI), along with UNDP [United Nations Development Programme] and the Swedish International Development Cooperation Agency (SIDA) in Ethiopia. It will also contribute to the capacity growth of the Ministry of Environment, Forest and Climate Change (MEFCC) and to the Ethiopian government’s overall capacity in the forest sector at all levels in order to spearhead the Climate Resilient Green Economy (CRGE) strategy and the Growth and Transformation plan (GTP) targets.”⁹²

“MoFEC [Ministry of Finance and Economic Cooperation], MEFCC, UNDP, the Norwegian Embassy in Addis Ababa and the Norwegian International Climate and Forest Initiative (NICFI) had a common understanding that a midterm review (MTR) of the Programme would be undertaken. The Review shall contribute to the quality and delivery of the remaining phase of the Programme. The review covers both the

⁹² Ibid. p.5.

CRGE and the UNDP components of the Programme. The MTR also highlights successes and challenges of the Programme, and where relevant, provides specific recommendations for improvement. A review is also considered particularly relevant because the activities of the Programme will be significantly scaled up in the afforestation/reforestation component of the REDD+ Investment Plan.”⁹³

The specific objectives are briefly but clearly described, and state that progress should be assessed against programme objectives. The context is described with an explanation of policies, Ethiopia’s recognition of the importance of forestry sector, and the involvement of donors. The programme similarly is described well, including a log frame, general structure, organisation and donor contributions.

There is no explicit discussion of a conceptual framework or approach; rather than presenting an approach or conceptual framework, the review lists methods used to collect data. However, the methodology, evaluation criteria and connected questions are described in an “evaluation framework matrix”. This has sufficient structure and level of detail

⁹³ Ibid. p.5.

to provide a good understanding of the overall design and approach to respond to the evaluation questions. The matrix clearly distinguishes between evaluation criteria, evaluation questions (a total of 31), data collection methods and sources, and data analysis methods. The report contains a very clear description of methods and how they will be used for data collection regarding each of the OECD DAC evaluation criteria, as illustrated by the following description of the desk review:

“The main purpose of the desk review was (i) to collect key background information on the Programme to inform the MTR team and (ii) to summarise the reported Programme outputs and emerging outcomes for field study verification. [...] The desk review has consolidated the background information provided on the Programme during the inception phase as well as the data collected during the field mission.”⁹⁴

The principles for selection of field sites and interviewees are well-described, providing clear and transparent information about the grounds for selection as well as representativity and approaches to compensate for limitations:

⁹⁴ Ibid. p. 6.



“At least one beneficiary community was visited in each woreda, and within the community, the MTR team held a minimum of two focus group discussions per site. The MTR team separated male and female beneficiaries for the purpose of conducting focus group discussions (FGD) and asked for FGDs to be organised in ways which reflect the make-up of the community (i.e. including female headed households, older and younger beneficiaries and wealthier and poorer beneficiaries). Before going to the field the team talked to relevant field project officials about the sample frame to be used, ideal FGD numbers (usually not too large a group works best), times of day and locations (women need a safe and private place for discussion, for instance). Representativeness depended upon the availability of a suitable sample frame. In the time available the data collected may best be seen as suggestive and indicative but no more. [...] Since the three woredas within SNNP [Southern Nations, Nationalities, and People] (Mirab Abaya, Sodo and Limu) were not visited, a tailored checklist questionnaire covering the key questions for investigation was sent to the SNNP relevant contacts to fill out.”⁹⁵

⁹⁵ Ibid. p. 6.

The report clearly states that the selection of field sites was logistically driven to ensure that the most woredas (districts) and regional offices could be visited in the limited time available. The report also describes the process of documenting, triangulating and analysing interview data:

“FGDs were run by the MTR team, who are skilled in rural facilitation and communication techniques. Probing, evidence checking and consensus building techniques were all used for in-group triangulation of data. Note-takers familiar with the local language recorded the fine grain of responses and typed it up in English each evening, so that the richest possible dataset was gathered in the time available from field interactions. The MTR team and note takers reviewed their joint findings at the end of each day, so that lessons learned could be carried over to the next day.”⁹⁶

The evaluation makes use of existing monitoring and evaluation data on activities and outputs and discusses its quality. The programme’s recording and reporting are discussed in several parts of the report, with special focus on the complicated demands and problems of financial reporting.

⁹⁶ Ibid. p. 7.

The main methodological weaknesses of the mid-term review are that ethical issues are not mentioned and that reliability and validity are not explicitly discussed. However, limitations are discussed relating to sampling and insufficient data on certain items, together with consequences.

The chapter on findings is organised according to the OECD DAC evaluation criteria and presents findings for each evaluation question in turn, clearly backed up by evidence. The findings and conclusions emphasise the importance of both community involvement (which is linked to acknowledging land rights) and commitment at the highest level. The slow response by the private sector was cited as an unexpected problem.

Relevance is assessed mainly in relation to three evaluation questions that refer to both Norwegian and Ethiopian forestry-related policies and to capacity development. It is reported that the programme was making major contributions to the partnership agreement between Ethiopia and Norway and hence to the Norwegian International Climate and Forest Initiative.

Effectiveness is assessed in relation to the nine evaluation questions covering both progress of



reforestation and capacity development of national and local authorities and to some extent local communities. The report finds that monitoring and reporting is insufficiently systematic and is poorly harmonised among donors. Although local-level staff spend a great deal of time on monitoring and reporting, the reporting that reaches the donor is inadequate. Risks are discussed in a separate section on risk management. A range of aspects are discussed, including organisation, limitations in capacity development at different levels, and environmental risks such as unexpected frost at higher terrain.

Efficiency is assessed in relation to six evaluation questions and includes both cost-effectiveness and organisation. Financial risks are also assessed, and the report notes that:

“the external review of the financial information provided found financial reporting limited and difficult to align with overall Programme spending and agreements. As far as is known, there is no consolidated or consistent format (accounting/ currency/period) for the programme financial reporting since the start of the programme. It has also been difficult to analyse the audit reports and financial

*statements as these have not been provided for the complete duration of the programme”.*⁹⁷

Sustainability is assessed in relation to three evaluation questions pertaining to expected lasting effects of the programme, ownership and the likelihood that activities will continue after donor support is terminated. Impact assessment was not included in the terms of reference. The review found that:

*“A sense of ownership of programme components is well-embedded in MEFCC, and the Government takes a great interest in the Programme and its achievements and progress. Local feelings of ownership are also strong at Woreda level, and very strong at Community level.”*⁹⁸

The report corresponds well to the terms of reference. All evaluation questions are extensively answered in responses that are well-grounded in clear sources. The intervention logic and results framework are presented and discussed in detail. Findings are well-presented in relation to the evaluation questions and have

⁹⁷ Ibid. p. 48.

⁹⁸ Ibid. p. 48.

explicit sources. Information from different sources is cited and insufficient data on certain items and the consequences are discussed. There is good analysis on the different outputs and outcomes and their relation to activities, including a discussion on continuing the intervention and likely future changes caused by the programme.

Conclusions also are well-presented under their respective criterion headings, including risk management, and there are clear references to relevant findings in the preceding findings chapter. Recommendations are organised and targeted according to both organisational level (national or woreda) and issues (coordination with other efforts, simplification of administration, gender, farmer support, technical forestry issues, and the role of the Norwegian Embassy). All recommendations are based on findings and conclusions by the review and seem actionable. The report does not identify lessons learned, although the terms of reference asked for these. However, some conclusions and in particular several recommendations comprise lessons learned.

Gender equality and women’s rights issues are discussed in relation to involvement in decision-making, execution of the programme and income-



generating activities. During field visits by the review team, separate focus groups with women were formed to ensure better responses to certain questions. Disaggregated data are used when available. The programme focuses on climate and environmental issues and these are thoroughly considered in the mid-term review. However, one of the recommendations also suggests that:

“... capacity building for financial management is as essential as forestry training. The need is for simple-to-follow report outlines with headings and table templates, which can be used each time a report is submitted to the donor. A standardised progress reporting template by UNDP and MEFCC would assist programme monitoring and management. Similarly, it may be necessary to generate a simpler financial data manual for the programmes which the Ministry can use for collation and analysis”.⁹⁹

⁹⁹ Ibid. p. 51.



List of Annexes

Annex 1: Terms of References

Annex 2: Data Collection Tools

Annex 3: Presentation of Data

Annex 4: Best Practise Evaluations

Annex 5: Methodology

Annex 5.1: Evaluation Matrix

Annex 5.2: Risks and Limitations

Annex 5.3: Request for Reports from Reviews and Decentralised Evaluations

Annex 6: Decentralised Evaluations Included in the Assessment

Annex 6.1: List of Quality Assessed Decentralised Evaluations

Annex 6.2: Descriptive Statistics of Quality Assessed Decentralised Evaluations

Annex 7: Profile of the Assessment Team

Annexes 5 to 7 can be found in Part II of the report on <https://www.norad.no/evaluation>



List of Tables and Figures

Table 1: Scoring Scale Used in the Quality Assessment Process	12
Table 2: Number of Decentralised Evaluations Included in the Annual Quality Assessment	13
Table 3: Average Scores, Report Quality Criteria	66
Table 4: Average Scores, Terms of Reference Quality Criteria	69
Figure 1: The Scoring Process	14
Figure 2: Distribution of Scores for Each Evaluation Report	18
Figure 3: Average Scores: Report Quality Criteria	20
Figure 4: Distribution of Scores: Summary, Style and Structure	22
Figure 5: Distribution of Scores: Evaluation Purpose, Objectives and Scope	24
Figure 6: Distribution of Scores: Methodology	27
Figure 7: Distribution of Scores: Application of Evaluation Criteria	33
Figure 8: Distribution of Scores: Analysis, Findings etc.	37
Figure 9: Distribution of Scores: Cross-cutting Issues	40
Figure 10: Distribution of Scores for Each Terms of Reference	42
Figure 11: Average Scores for Terms of Reference Quality Criteria	43
Figure 12: Distribution of Scores: Terms of Reference Quality Criteria	45
Figure 13: Distribution of Scores per Report	71
Figure 14: Distribution of Scores per Report Quality Criterion	72
Figure 15: Comparison of Average Scores for Reports Referred to as ‘Evaluations’ and ‘Reviews’	73
Figure 16: Distribution of Scores per Terms of References	74
Figure 17: Example of the Layout in the Findings Chapter of UN Women Malawi (2019)	80



Acronyms and Abbreviations

CRGE	Climate Resilient Green Economy
DAC	Development Assistance Committee
FGD	Focus Group Discussions
MDG	Millennium Development Goal
MEFCC	Ministry of Environment, Forest and Climate Change
MTR	Midterm Review
Norad	Norwegian Agency for Development Cooperation
OECD	Organisation for Economic Co-operation and Development
REDD+	United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation
UN	United Nations
UNDP	United Nations Development Programme
UNIDO	United Nations Industrial Development Organization



EVALUATION DEPARTMENT



Norwegian Agency for
Development Cooperation

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

Cover photo Ken Opprann
ISBN 978-82-8369-046-0
October 2020