EVALUATION DEPARTMENT

REPORT 1/2017

Norad

The Quality of Reviews and Decentralised Evaluations in Norwegian Development Cooperation

CONTENT

Commissioned by
the Evaluation Department

Carried out by
Itad in association with Chr. Michelsen Institute (CMI)

Written by
Nick Chapman (team leader, Itad), Rob Lloyd (Itad),
Espen Villanger (CMI) and Greg Gleed (Itad).

JANUARY 2017

Norad
Norwegian Agency for
Development Cooperation
www.norad.no
post-eval@norad.no

January 2017
Photo: Ken Opprann (cover)
ISBN: 978-82-7548-843-3

Foreword

This evaluation was initiated based on indications from several studies that the quality of reviews and desentralised evaluations in Norwegian development cooperation was variable. The annual government budget proposal for 2016 announced the introduction of stricter requirements to undertake reviews and evaluations in the project cycle.

With increasing demand for evaluation to document the results of Norwegian development cooperation, there is need for more information about the quality of these evaluations, as well as a better understanding of factors contributing to quality and use of evaluation findings, conclusions and recommendations.

The purpose of this evaluation has been to contribute to good quality reviews and decentralised evaluations in Norwegian development cooperation.

The evaluation was carried out by the British consultancy company Itad Ltd. in collaboration with the Chr. Michelsen Institute, Norway. We thank the team for a job well done.

Oslo, January 2017

Per Øyvind Bastøe
Director, Evaluation Department

Acknowledgements

The evaluation team would like to express their warm thanks to the various grant managers, aid officials and consultants as well as the Evaluation Department in Norad, and who gave their time to help conduct this meta-evaluation.

DISCLAIMER

The views expressed in this report are those of the evaluators. They do not represent those of the Norwegian aid administration or of any of the individuals and organisations referred to in the report.

THE EVALUATION TEAM AND THEIR ROLES

The evaluation team consisted of four persons: three from Itad and one from CMI. Team leader, Nick Chapman (Itad), team members Rob Lloyd (Itad) Espen Villanger (CMI) and Greg Gleed (Itad). The team leader guided the work of the team, and ensured all team members contributed to the work as required and according to their experience and skills. Our Norwegian speaker from CMI accessed Norad and MFA archives to extract relevant supporting documentation available in Norwegian for the quality review and case studies.

Executive summary

INTRODUCTION

This meta-evaluation examined a set of reviews and decentralised evaluations (hereafter termed 'reviews') commissioned by various arms of the Norwegian aid administration in 2014.[1]

PURPOSE

The overall purpose of this evaluation is to contribute to good quality reviews in Norwegian development cooperation. The intended users are the Ministry of Foreign Affairs, Norwegian embassies managing official development assistance (ODA) funds and Norad.

There were three specific objectives: (1) assess the quality of reviews, (2) examine the use of the review findings and (3) identify factors contributing to quality and use.

1 The Ministry of Foreign Affairs and Norad's Grant Management Manual defines a review as 'a thorough assessment with focus on the implementation and follow-up of plans', which may be undertaken under way (mid-term review) or after finalisation to assess the effect of the programme/project (end review).

FINDINGS

1. There was considerable variation found in the different quality areas for the reviews and their accompanying terms of reference.

The highest quality scores for reviews were found in relation to stating the purpose of the review, defining the object to be reviewed, answering the questions posed in the TOR, and making useful recommendations. Lowest quality scores were found in the description of the methods to be used in the review, dealing with ethical issues and examining the programme's logic. The highest quality scores for TORs were found in criteria concerned with the review rationale and purpose, the specific objectives and scope, and the description of the review process and deliverables. Criteria which were scored as lower quality in TORs included the context, review criteria, cross-cutting themes, ethics and limitations.

2. Reviews were generally considered to be timely and to present relevant and realistic recommendations.

Timeliness was linked to the eventual use of a review and whether it influenced upcoming decisions. Recommendations were considered relevant and realistic when they fed directly into the needs of the users, both for the Norwegian aid administration and the grant recipient. The ability of reviews to deliver concrete recommendations was a key factor in the use of reviews.

3. There was a high level of use of reviews.

Reviews were well used by the unit responsible for managing the grant for these interventions. Reviews were most often used in *instrumental* ways (for management of and decisions related to the intervention being evaluated). *Conceptual* use (for wider learning or policy beyond the specific intervention being evaluated) was limited. *Symbolic* use (where a review was used to justify a decision or as a routine requirement in closing an intervention) was higher than

would have been expected. Given that one would expect under-reporting of symbolic use, we consider it to be significant that a relatively high number of respondents reported such use.

Four main factors determined use of reviews:

1. Formulation of high quality TORs and the delivery against them, influenced by stake-holder engagement from the beginning and ensuring clarity of purpose of the review task

2. The production of realistic and actionable recommendations

3. Planning and delivering reviews in a consultative way

4. Ensuring a review was completed at the right time to feed into a decision

4. The quality score of the review TOR, the level of resources allocated to a review and the calibre of the review team were the main factors contributing to quality scores of reviews.

There was a significant and positive statistical relationship between the quality scores of the TORs and the quality scores of the final review reports.

There was evidence that the reviews for projects with larger budgets and reviews that had more days allocated for the work received higher quality scores. This was consistent with a number of other studies on evaluation quality which have also found that resources (budget and days) and quality are linked.

The calibre of the review team, including appropriate evaluation expertise, and knowledge of the context, subject matter and the project were important determinants of reviews with higher quality scores.

5. A majority of reviews were not based on data and analyses that were likely to produce credible information.

The quality of the findings presented in many reviews was unsatisfactory due to weaknesses in the methodology and analysis. We found that over 65% of the reviews did not contain sound methodological underpinnings that would support or produce credible findings. Often reports had only a very limited discussion of methodology, and some had none at all. The quality criteria related to methodology (such as data collection, analysis, limitations and ethics) had the lowest ratings across the sample, and this weakened the possibility of the reviews producing credible findings.

6. There was limited evidence to suggest that staff in the Norwegian aid administration believe that robust methodology is important to quality.

Staff in the Norwegian aid administration gave limited consideration to methodological rigour in assessing the quality of reviews. Instead,

they focused on actionable insights, conclusions and recommendations. The challenge with this approach is that grant managers and project implementers may have taken actions based on evidence that was of poor quality.

7. Grant managers did not have access to the necessary tools to conduct reviews.

The ability of grant managers to conduct reviews was limited by a gap in the provision of technical guidance and a lack of access to up-to-date information for grant managers to use in making management decisions. The Grant Management Manual is the only official source of guidance for commissioning and managing reviews, and information regarding planned and completed reviews is not regularly updated by all grant-managing units.

8. There were few lessons of broader applicability that were generated by the reviews and the reviews gave limited attention to the wider political, environmental and social aspects that the aid project was embedded in.

We found only seven reviews which provided significant learning or lessons of wider applicability to inform Norway's aid agenda. Where lessons of broader relevance did exist, they tended to relate to programme design and delivery issues. The fact that attention to context was limited and that very few reviews provided lessons learned, limits the opportunities for review findings to contribute to the design and implementation of Norwegian aid interventions and policies beyond the aid project reviewed.

CONCLUSIONS

We conclude that decisions about Norwegian aid projects are being taken based on review findings and recommendations that are not always grounded on sound evidence. While the limitations of this quality review need to be acknowledged (see below), the evaluation suggests that a majority (over 65%) of the reviews conducted on Norwegian aid programmes are unlikely to contain sound methodological underpinnings that would support or produce credible findings.

However, ***reviews are important management tools for the units responsible for the grants.*** Based on the evaluation's evidence, review reports were highly used, and considered very useful in the aid administration, with a focus mostly on instrumental use to inform an ongoing or planned programme. Nevertheless, whether reviews are used for these purposes or for conceptual use, there should be a sound information base that is used in an analytical way to draw conclusions.

METHODOLOGY

The evaluation methodology had four components: (1) an email survey of grant managers, (2) a quality assessment of 60 reviews and associated TORs conducted in 2014, (3) case studies from five reviews out of the 60, and (4) an online survey of staff from Ministry of Foreign Affairs (MFA), Norad and Norwegian embassies. Together, these components gathered evidence to address the six evaluation questions.

LIMITATIONS

There were three main limitations to this evaluation.

Approach to defining quality focused on adherence to OECD-DAC standards: Given the time frame and resources available for the evaluation, the ability to explore the merit of and the validity of review findings was limited.

Size and choice of the review and survey samples: Examining cases from only one year, 2014, brought the risk that results may not reflect the level of quality in other years. A sample of 60 reviews was likely to be sufficient to gain an overall picture of quality in 2014, but further disaggregation led to small sub-samples that could not accurately gauge the significance of explanatory factors. The online survey had a low number of responses: 34 responses out of 120 contacted, 28%. Although the percentage of responses was not untypical of such surveys, the response sample size was smaller than we would have liked.

Difficulty gaining access to documents: To mitigate this, the email survey to grant managers allowed for surfacing of additional documents, as well as for additional data on evaluation budget and team composition. In the sampling approach, selected cases were replaced that lacked documentation with other cases where more complete documentation existed. Nevertheless, the data obtained for analysis were limited (particularly TORs and inception reports) despite an earlier mapping study, support from the Evaluation Department in NORAD and support from MFA to conduct an archive search of its records.

1. Introduction and purpose

This meta-evaluation examines a set of reviews and decentralised evaluations[2] conducted in 2014 by various arms of the Norwegian aid administration. It draws on a Mapping Study in 2015 that assembled a set of 274 review reports covering the period 2012–15.[3] These reports were conducted by different entities of the Norwegian aid administration including the Norwegian Agency for Development Cooperation (Norad), the Ministry of Foreign Affairs (MFA) and Norwegian embassies abroad.[4]

The MFA and Norad's Grant Management Manual (GMM) defines a review as *'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). Guidance for why, when and how to undertake reviews is given in the GMM and further requirements are specified in the rules for each grant scheme.[5]

As stated in the TOR, the *'overall purpose of this evaluation is to contribute to good quality reviews and decentralised evaluations in Norwegian development cooperation'*. The main intended users are the Ministry of Foreign Affairs, Norwegian embassies managing official development assistance (ODA) funds, Norad and other parts of the aid administration.

The evaluation will serve as input into a discussion on the organisation and management of decentralised evaluation and reviews (hereafter termed 'reviews') in the Norwegian development administration.

The specific objectives are to:

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

2. Examine the use of review findings, conclusions and recommendations; and

3. Identify factors contributing to quality and use of reviews and decentralised evaluations in Norwegian development cooperation.

These objectives are addressed through a set of six evaluation questions:

1. What are the main strengths and weaknesses of reviews and decentralised evaluations of Norwegian development cooperation?

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

2 Decentralised evaluations are evaluations commissioned by the unit responsible for grant management (embassies, MFA, Norad), implementing partners/grant recipients and other agencies/co-sponsors. These are normally referred to as reviews, while the term evaluation is for larger studies with broader scope (TOR, p. 1).

3 NORAD (2015) Study of Reviews and Evaluations in Norwegian Development Cooperation – Mapping. Oslo: The Evaluation Department. Report no. 11.

4 The scope excludes studies undertaken by the Evaluation Department of Norad, appraisals, studies and forensic audits, and review and evaluation reports commissioned by NGOs receiving Norwegian grants. It also excludes organisational reviews and thematic studies.

5 Grant Management Manual, Management of Grants by the Ministry of Foreign Affairs and Norad, 2013.

3. From the perspective of stakeholders, to what extent are reviews timely, and present relevant and realistic recommendations?

4. To what extent have review findings, conclusions and recommendations been used by the unit responsible for managing the grant to the intervention that has undergone review?

5. What are the main factors contributing to quality and use of reviews and decentralised evaluations?

6. To what extent do reports present any general lessons learned with relevance beyond the intervention under review?

2. Context

Reviews form a key part of the evidence base for documenting the results of Norwegian development cooperation. These reviews and evaluations are central monitoring and assessment tools of the internal grant management system, which guides how the Norwegian aid administration manages the finances disbursed through the Norwegian aid budget. This system generates the bulk of grant results data. Although reviews are also used for purposes other than results documentation, such as for organisational assessments and producing thematic overviews, these fall outside the scope of this evaluation (see TOR in Annex 1).

In the Norwegian grant management system, the recipients of the grants, who are external to the aid administration, implement all projects. The grant recipient is responsible for measuring and reporting the results of the grant through periodic progress reports and end reports, which typically consist of self-evaluation in a narrative form. Reviews are intended to play an important role in complementing the results documentation with more thorough and independent inquiries.

Reviews are defined by the GMM as thorough assessments with a focus on implementation and performance in relation to plans and goals (GMM, pp. 9, 66). Mid-term reviews focus on mid-way implementation and follow-up of plans while end reviews address the final stage and the documentation of the effects of the project or programme, including assessing outcomes and impacts (GMM, p. 66). The reviews are a formalised follow-up of the project or programme commissioned by the Norwegian aid administration, and the GMM provides a clear description of the expected content of the reviews.

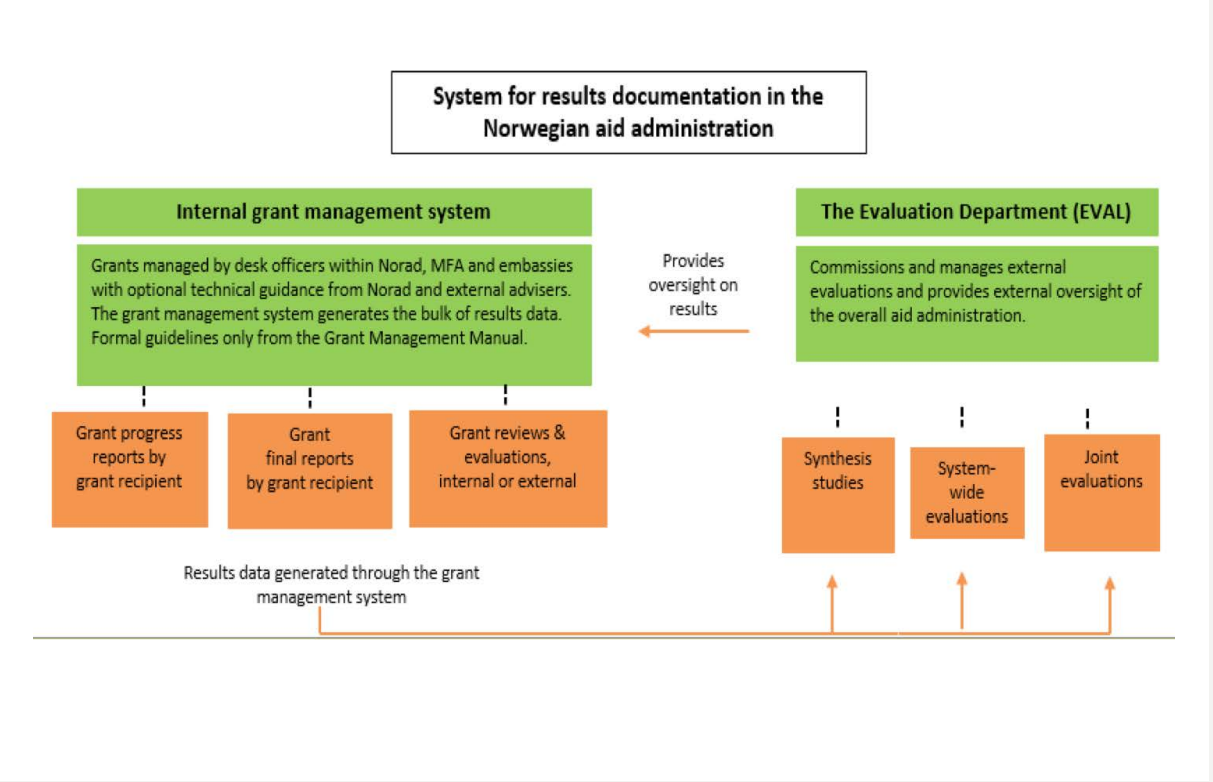
The grant manager has the central role in the results documentation and is responsible for the collection, assessment and follow-up of the grant recipient's information on results throughout the grant cycle (GMM, p. 81). The grant manager would therefore also be one key user of the reviews. Moreover, since the MFA and Norad have the overall responsibility for the collection and assessment of the results of the grants, there are many potential users in these hierarchies including the Evaluation Department (GMM, p. 90).

Figure 1 (next page) shows the role of the grant reviews in relation to the overall system for results documentation in the Norwegian aid administration with its two main components: the internal grant management system, and the Evaluation Department. This underlines the central use of reviews in the documentation of the results as it feeds into the Evaluation Department's work to provide an overview of the results of Norwegian development cooperation.

As indicated in Figure 1, the GMM is the only formal source governing reviews in the Norwegian aid management system where a grant manager can seek guidance on conducting review work. There are no handbooks, standards, templates or guidance notes, nor any help desk or external quality assurance function to aid grant managers in commissioning and managing reviews.[6] However, the GMM provides a short description of why, when and how to undertake reviews (GMM, p. 66), and explicitly refers to a guide in the manual on management of results and risks (GMM, pp. 81–92).

In addition, in the period covered by our evaluation (2014), any requirements for reviews would be specified in the rules for each grant scheme, but that would usually be a minimal description of what was needed and whether it was mandatory for projects to be subjected

FIGURE 1: THE FORMAL ROLE OF REVIEWS IN DOCUMENTING RESULTS OF NORWEGIAN AID



Source: Norad 'Can we demonstrate the difference Norwegian aid makes': The Evaluation Department, Report 1, Oslo 2014

6 This is examined further in the report 'Can We Demonstrate the Difference that Norwegian Aid Makes? Evaluation of Results Measurement and How This Can be Improved', Oslo: Norad, 2014.

to reviews.[7] Unless it was stated explicitly as mandatory, each grant manager had the responsibility to decide whether a grant should be reviewed or not, based on an assessment of the risk and significance of the project. This changed in 2016 when the MFA introduced new grant scheme rules where reviews became mandatory for programme/project agreements of a duration of over two years, and for agreements above a certain financial threshold, depending on the grant scheme.

The GMM stipulates that the following factors *should* be addressed in a review:

- results in relation to the goal hierarchy (results framework) implementation plans and budgets
- efficiency and effectiveness
- risks, and
- the capacity of the grant recipient and the models and methods employed in the project or programme.

The focus of reviews should thus be on operational aspects and factors influencing implementation and on whether the project inputs and activities achieve their objectives and hence the project's degree of effectiveness. In addition, efficiency – or value for money – is of prime interest, while the other OECD-DAC criteria 'relevance' and 'sustainability' are not in focus, although capacity and risk assessments may be elements of these.

Furthermore, the GMM states that the following factors *may* be included:

- the effect of the project or programme in relation to external factors,
- the benefit achieved through changes in the operating conditions, and
- the need and potential for reducing risk.

The GMM also requires that cross-cutting issues in Norwegian development cooperation should be taken into consideration in all interventions. In the period under evaluation, these were women's rights and gender equality; climate and environment; and anti-corruption. The GMM does not explicitly state that reviews should take these issues into account. However, cross-cutting issues could be seen as part of risk management, and as such, among the aspects to be covered by a review.

In preparation for this evaluation, the Evaluation Department commissioned a Mapping Study[8] in order to get a better overview of the extent of reviews in Norwegian development cooperation. The exact number of reviews undertaken per year is not known, but the Mapping Study identified 235 reviews in the period January 2012 – May 2015, 60–70 per year. Of these, only 60 % had TORs attached to the reports.

As noted in this evaluation's TOR, 'responsibility for tracking and collecting reviews done throughout the aid administration is not clear. The grant management system (PTA) of the

7 For example, the grant scheme rules for Peace, reconciliation and democracy 2014 stated the following (on p. 5) about evaluation of the grant scheme: 'Norad's Evaluation Department is responsible for planning and conducting independent evaluations of activities funded under the Norwegian development budget. Evaluations may be carried out of the whole grant scheme or of parts of it, or of cross-cutting themes, common objectives or countries that are covered by several grant schemes. The Ministry will also participate in joint evaluations and reviews with other donor organisations, the UN system and partner countries. In addition, the unit responsible may also initiate independent evaluations.'

8 Mapping Study, op. cit.

Norwegian aid administration has a report function in place to track reviews, and the GMM has a requirement to register planned and completed reviews in PTA. This represents a potential source of credible information about the extent of reviews and evaluations. However, it is not kept updated by all grant-managing units. Therefore, it is not known how much of the annual development cooperation budget is subject to a review, and whether it is the most significant programmes that are reviwed.'

The TOR also notes that 'Reviews may be published at www.norad.no as part of the report series Norad Collected Reviews, though this is not a requirement in the GMM or the grant scheme rules'.[9]

9 New grant scheme rules, as of February 2016, stipulate that reports be submitted to the Evaluation Portal, managed by the Norwegian Government Agency for Financial Management, in which all evaluations by government agencies should be registered. This should ensure an overview of reviews and evaluations in the future.

3. Methodology

The evaluation had four components: (1) an email survey; (2) a quality assessment of 60 reviews; (3) case studies of five reviews; and (4) an online staff survey. These were designed to address the six evaluation questions (EQs) set out in Chapter 1. How each component and its analysis were designed to answer each question are set out in Table 1 (next page).

The following sections detail the evaluation's approach to measuring evaluation quality and use. Annex 3 describes each component of the methodology in detail.

3.1 DEFINING QUALITY

The literature on measuring evaluation quality is broad, and there are various interpretations of how to judge quality – from adherence to international standards and norms to the usefulness of the content and whether new ideas or insights are produced.[10] Some approaches combine ethical standards together with quality.[11] Others explore the perspectives of the evaluator, the evaluand and the commissioner, and how this affects quality and use.[12]

While recognising these perspectives, quality in this study was judged by adherence to the internationally established set of standards produced by the OECD-DAC.[13] This had the strong advantage that the assessment was based on a well-known and transparent set of judgement criteria shared by the global evaluation community. It also was most suited to a meta-evaluation such as this one, where a large set of reviews or evaluations are to be assessed and compared for quality. In addition to the OECD-DAC standards, Norad was also interested to see how far the cross-cutting themes as defined in the GMM were also covered.

The quality of a review report was therefore defined in this evaluation to be the degree to which (1) it did or did not apply a set of quality areas as defined by the OECD-DAC evaluation standards and, in addition, Norad's cross-cutting themes; and (2) applied these quality areas in a way that provided or generated trustworthy information.

In operationalising this definition, 32 quality criteria were defined for the review report and 15 quality criteria for the TORs. For the reviews, these were divided into five quality areas:

- Summary, style and structure
- Review purpose, objectives, object and scope
- Methodology
- Application of selected OECD-DAC evaluation criteria (such as relevance, efficiency, impact)
- Analysis, data, findings, conclusions, lessons and recommendations

10 T. Schwandt, Defining 'Quality' in Evaluation. *Evaluation and Program Planning*, 13(2): 1990.

11 USAID, *Meta-evaluation of Quality and Coverage of USAID Evaluations 2009–12*, prepared by Management Systems International. UNDP, Annual Report on Evaluation, Independent Evaluation Office, UNDP, 2013.

12 B. de Laast, Evaluator, Evaluand, Evaluation Commissioner, a Tricky Triangle, Ch. 2 in Loud and Mayne (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*, Thousand Oaks, CA: Sage, 2014.

13 Organisation for Economic Cooperation and Development, Quality Standards for development evaluation, DAC Guidelines and Reference Series, Development Assistance Committee, 2010.

TABLE 1: EVALUATION OBJECTIVES, QUESTIONS, METHODS AND ANALYSIS

Evaluation objective 1: Assess the quality of reviews and decentralised evaluations of Norwegian development cooperation		
Evaluation question	**Method**	**Analytical process**
1. What are the main strengths and weaknesses of reviews and decentralised evaluations of Norwegian development cooperation?	Quality assessment of 60 review / evaluation reports and their TORs (51 were available) using standardised quality template Online Survey of MFA, Embassy, Norad	Analysis of scores across quality areas to reveal common strengths and weaknesses Use of statistical methods to examine patterns and correlations within data and explore effect of variables such as evaluation type, budget etc. on quality. Coding of qualitative data recorded as part of the quality review of evaluations, to illuminate quantitative findings and provide deeper understanding of quality areas
2. To what extent are the reviews and decentralised evaluations based on data, methods and analyses that are likely to produce credible information about the programmes and their outcomes?		Analysis of scores across specific quality areas related to methodology. These include quality and appropriateness of: data collection tools, data sources, sampling and data analysis approach
3. To what extent do reports present any general lessons learned with relevance beyond the interventions under review?*		Analysis of scores across quality areas related to nature of the lesson identified in the review and assessment of the extent to which they are context specific or transferable
Evaluation objective 2: Examine the use of review findings, conclusions and recommendations		
4. From the perspective of stakeholders, to what extent are reviews timely, and present relevant and realistic recommendations?	Case studies of five review processes Email Survey of 73 grant managers Online staff survey of MFA, Embassy, Norad	Case studies of five reviews provide in-depth analysis of whether and why a sample of reviews were timely and presented relevant and realistic recommendations The survey of grant managers responsible for the 2014 pool of reviews gathers information on the perceived standard and utility of each review from the point of view of the commissioner, as well as additional information on budget, team composition and relevant documents The online staff survey provides more general data on timeliness and relevance Together these data provide insights into review use from the perspective of different stakeholders the extent to which reviews are timely and present relevant / realistic recommendations
5. To what extent have review findings, conclusions and recommendations been used by the unit responsible for managing the grant to the intervention that have undergone review?		

*This question was placed sixth in the TOR, but we have moved it to third, as it will be mainly assessed together with EQs 1 and 2 from the quality review.

→

Evaluation objective 3: Identify factors contributing to quality and use of reviews and decentralised evaluations in Norwegian development cooperation		
Evaluation question	**Method**	**Analytical process**
What are the main factors contributing to quality and the use of reviews and decentralised evaluations?	Quality assessment of 60 review reports and TORs using standardised quality template Case studies of five review processes Online staff survey of MFA, Embassy, Norad	Analysis of scores across quality areas to reveal common strengths and weaknesses. Use of statistical methods to examine patterns and correlations within data and explore effect of variables such as review type, budget etc. on quality. Coding of qualitative data recorded as part of the quality review of reviews, to illuminate quantitative findings and provide deeper understanding of quality areas Triangulation with data emerging from the staff survey and case studies to see if patterns emerging from the quality assessment align with case study and survey findings

For the TORs, these were divided into three quality areas:

- Review purpose and objectives
- Review process
- Overarching and cross-cutting themes

Each quality criteria specified how it should have been applied and/or what information should be generated. Annex 3 contains the template with the detailed definitions of each quality area and critiera. Appendix 1 in Annex 6 provides an overview of how each OECD-DAC evaluation standard was mapped against the review and TOR templates as well as the other instruments.[14]

Quality ratings

The use of quantitative ratings linked to qualitative descriptions has been found in previous meta-evaluations[15] to provide a number of advantages: (1) the qualitative descriptions embrace and acknowledge that quality in evaluation requires a degree of judgement, and this is made as transparent as possible; (2) the inclusion of ordinal ratings allows for quantitative analysis to be undertaken in order to find patterns; and (3) the use of a four-point rating scale forces the reviewer to rate each quality area in the TOR or review as satisfactory or not, and allows for a degree of disaggregation in the analysis.

Each quality criteria was therefore awarded a rating according to a four-point scale centred on 'satisfactory quality' as described in the OECD-DAC standards. Scores of 4 or 3 denote the reviewer having confidence in the review, meeting quality standards to a good or adequate level. Scores of 4 were reserved for the cases where the review report delivered

14 Given the limitations of the evidence available to the evaluation team, there were aspects of the quality of the evaluation process and/or outcomes that could not be assessed (for example the participation of stakeholders, the tone of the relationship between the team and the management, any instances of capacity development). It was not strictly possible, therefore, to cover all the OECD-DAC standards based on the two documents available.

15 For example, UNICEF, The Global Evaluation Report Oversight System (GEROS) 2010–15, 2015; UN Women, Global Evaluation Reports Assessment and Analysis System (GERAAS), 2014; Global Affairs Canada (2016) *Meta-evaluation of Global Affairs Canada's Decentralised Evaluations*: FY 2009/10–13/14, due for publication late 2016, Ottawa.

in a complete way on the described quality definition (i.e. covering all the specified requirements or no substantial shortcomings). Scores of 3 were given when there were some shortcomings but where the overall quality criteria was still assessed to be satisfactory. Scores of 2 or 1 denote that the reviewer assessed the quality of the criteria to be unsatisfactory, assigning a rating of either less than adequate or poor. Scores of 1 were given when a quality criteria was not applied, or when the quality of what was delivered when applying the criteria was clearly poor. If the review contained some elements of good application or information about the quality criteria, but where the overall delivery was not satisfactory, a score of 2 would be assigned. Table 2 shows the categorisation.

TABLE 2: RATINGS FOR QUALITY

Satisfactory		Less than satisfactory	
4	Good quality	2	Less than adequate quality
3	Adequate quality	1	Poor quality
Not relevant: The criterion was not included in the evaluation Not assessable: The criterion was included but it is not possible to assess quality because there is too little information			

The key strength of this approach is that the quality assessments can be compared across reviews for each review quality area. In that way, it is possible to identify the strengths and weaknesses in review quality at the detailed level, which is useful for understanding exactly where there is potential for improving review quality generally across the sample. In addition an overall average rating can be calculated to obtain a measure of general quality.

Overall ratings: Two overall quality ratings were calculated based on the mathematical average of 15 separate quality ratings for TORs and 32 separate quality ratings for the reviews.[16] The approach taken was to use an unweighted score – in other words each of the quality criteria had an equal weight in determining the overall average. This has the advantage that each quality criteria has an equal influence on the overall average score. But it also gives equal emphasis to very different criteria, some of which might be regarded as more important in determining quality or influencing use than others. For example, having a clear rationale or sound recommendations might be considered of higher merit than having a readable style or addressing ethics or cross-cutting themes well.

While the overall score was unweighted, it should be noted that certain quality areas contained more quality criteria than others and therefore those sections will have a greater influence on the overall rating. The most important quality dimensions in the review template (Quality area 2, 3, 4 and 5) together comprise 27 of the 32 quality areas and will *a priori* have the most influence on the

16 Cases where a 'not assessable' or 'not relevant' rating were given were not included in the overall average calculation.

overall rating. In this sense, we believe that the overall average rating calculated here represented a fair metric to assess the combined influence of the different quality areas in a single score.

In the three other components where respondents' views on review report quality were captured (email survey, case studies and online survey), the respondents applied their own definition. The potential divergence in definition between our team conducting the quality review and the grant managers and review users is discussed in Chapter 5.

3.2 DEFINING EVALUATION USE

Evaluations, as with any type of evidence vehicle, are used in a variety of ways in policymaking. Furthermore, estimating the potential value of such uses is not straightforward. For example, one analysis compared nine different approaches to measuring benefits arising from evaluation work and found that most methods were rather time consuming especially if they were to be applied before the evaluation itself was conducted.[17]

A recent overview of the literature in this area identified a broad typology of uses. This encompassed five categories of use: *instrumental* (use made of evaluation findings to directly improve a project); *enlightenment* (which can be conceptual or more short term to enhance knowledge about the type of intervention or issues under study, or reflective or longer term to explore wider or future strategies through an evaluation); *persuasive* (to build up support for an intervention or to criticise it); *process* (making use of the process of doing an evaluation to better understand the intervention); *symbolic* (an evaluation that fulfils a bureaucratic or programming requirement rather having its own intrinsic merit).[18] These potential categories would require quite deep investigation in order to produce sufficient evidence. It is also worth noting that it may sometimes be difficult to discern differences between these categories, and that they may overlap. For example, instrumental and symbolic use may equally appear to apply when an evaluation aligns with already planned changes.

Given the resource and time constraints of this evaluation, we used a simpler but still widely-used model for conceptualising evaluation use was adopted. This is the 'Stetler Model'.[19] It describes three types of evaluation use:

- *Instrumental:* Knowledge from an evaluation is used directly to inform an ongoing policy or programme;
- *Conceptual:* No direct action is taken as a result of the evaluation, but the knowledge from the evaluation influences people's general thinking around what works;
- *Symbolic:* When people use the mere existence of an evaluation, rather than

17 The Value of Evaluation: Tools for Budgeting and Valuing Evaluations, J. Barr (Itad), D. Rinnert (DFID), R. Lloyd (Itad), D. Dunne (DFID), A. Henttinen (DFID), Discussion Paper, August 2016, DFID.

18 These types are suggested by M. Loud and J. Mayne (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Unit*s, Thousand Oaks, CA: Sage, 2014, p. 3.

19 Stetler, C.B. (2010). Ch. 3: Stetler Model. In J. Rycroft-Malone and T. Bucknall (eds), *Models and Frameworks for Implementing Evidence-Based Practice: Linking Evidence to Action.* Evidence-based Practice Series. Oxford: Wiley-Blackwell.

its specific findings, to persuade or convince. A version of this – political/strategic use – is when an evaluation is used to justify or legitimate a policy or decision.

3.3 ETHICAL ISSUES

As part of designing and conducting any evaluation, it is important to ensure that appropriate ethical issues are addressed. These relate to matters such as individual confidentiality, their rights to privacy and respect, and to consultation and feedback during the process. In this case, all informants contacted during the email and online surveys and the five case studies were advised that their views would not be attributed directly. However, since there are only five case studies, and details about the cases will appear in the report, a respondent's identity could still be inferred by other people with detailed knowledge about the review.

In terms of consultation, we shared notes on interviews made during the case studies with interviewees and their comments were incorporated. The draft report was shared with a wide range of stakeholders for comment. Finally, the dissemination of the evaluation will use various means to provide feedback including an internal workshop and a public seminar.

4. Limitations

Using a normative approach to quality: The approach to quality in this evaluation largely focused on adherence to OECD-DAC standards and less on the substance of the findings/conclusions of the review. In this exercise, which was primarily a desk review of a set of TORs and review documents, there was little possibility to discern whether the report in fact presented evidence or conclusions that offered new and/or usable insights to the eventual users or beneficiaries. While our team brought its varied experiences to bear on the quality review, it was not constituted to bring expertise in all the sectors presented across the sample of reviews, and therefore did not make a judgement on the extent to which the findings were in fact new and offered realistic actions.

The need to address broader aspects of quality that include utility to the users in the aid system was nevertheless provided in the case studies and online survey components of this evaluation, where feedback from different involved actors helped reveal the merit of the report content through triangulation. At the same time, as far as the evaluation is concerned this is still at best second-hand experience, and even then the interlocutors (such as grant managers, sector specialists, consultant team leaders) are themselves not necessarily directly familiar with the empirical situation surrounding an intervention.

Sampling issues: The quality review sample of 60 was taken from the 74 reviews completed in 2014.[20] The sample is listed in Annex 5. Annex 6 examines the main characteristics of this sample of reviews against the universe as established by the Mapping Study in 2015. The conclusion from this analysis is that there was a reasonably close match between the sample and the universe in terms of key independent variables such as region, commissioner and type of review (Annex 6, Table 1).[21]

Out of the sample of 60 review reports, 51 (85%) were found to have TORs available for the quality review.[22] Tables 2, 3 and 4 in Annex 6 show the distribution of TORs by region, commissioner and type. The missing TORs were evenly spread across these categories, although significantly more were missing from embassy and partner commissioned reviews (6 out of 26 embassy reviews and 3 out 7 partner reviews did not have TORs).

The sample choice and size will affect the predictive power of the evaluation. The choice of a single year, 2014, for this exercise brought the risk that it may not reflect the level of quality in other years or that the drivers of quality may have operated in different ways in other years. This is examined more closely in Chapter 5. While the quality review sample size (of 60 reviews) was likely to be sufficient to gain an overall picture of quality, when analysing the different causal factors,

20 The Mapping Study identified 84 reports, but this was reduced by Norad Evaluation Department to 74 as organisational reviews and thematic studies were excluded.

21 In regional terms, the sample has a slight over-representation of cases from Africa South of the Sahara, and under-representation of cases from Americas and Global regions.

22 These TORs were those found in the documentation library prepared by the Mapping Study, supplemented by a further investigation by one team member in Oslo of Norad and MFA archives. Gaps are due in some cases to the practice in Embassies of holding non-electronic copies of such documentation.

any disaggregation will lead to quite small sub-samples that may have weak statistical power to gauge the true significance of those factors among the population. We therefore treated such findings with caution where differences were small, and used triangulation from the other data sources in the evaluation (such as the surveys and case studies) to build greater confidence in the results.

Response rate: The online survey response rate of 28% (34 responses out of 120 contacted) was lower than would have been desired, though not untypical of such online surveys.[23] The responses were predominantly from Norad and the embassies with fewer from the MFA sections (Annex 8), but the split between review commissioners and grant managers was quite even. Nevertheless, where findings in the evaluation were predominantly based on the results from this survey, they should be treated with some caution.[24] In many areas, however, the case study results align quite closely with the online survey findings.

Incomplete documentation: Gaining access to documents proved somewhat difficult despite the Mapping Study and support from Norad's Evaluation Department. To mitigate this, the email survey to grant managers allowed for additional documents to come to light, as well as for additional data on evaluation budget and team composition. Nevertheless, the data obtained here for analysis were limited.[25] We also deployed one member to work in Norad for one day to extract relevant documentation in an archive. In addition, support from MFA enabled a similar archive search of their records. In the sampling approach, selected cases that lacked documentation were also replaced with other cases where more complete documentation existed.

In assessing quality, the quality review component of the evaluation did not look at inception reports or tender documents. Although the email survey requested grant managers to provide these, the response was low, and only three inception reports were actually retrieved for analysis in the set of 60 reviews. This is potentially an important limitation as often the evaluation approach and methodology would likely be elaborated in these deliverables, rather than in the final review report. However, further analysis showed that the majority of the reviews in the 2014 sample *did not have an inception report* (see 5.2.4). Our judgement on the quality of review methodology therefore was based on the review reports alone. This was supplemented by the case study component which did examine all available supporting documentation for the five cases selected.

23 In a similar meta-evaluation survey for Global Affairs Canada, the response rate was around 10% (Global Affairs Canada (2016) Meta-evaluation of Global Affairs Canada's Decentralised Evaluations: FY 2009/10–13/14, 2016).

24 The report provides the specific response rate for each survey finding so that the user can judge the relative strength of evidence.

25 With only 16 cases where the evaluation budget was available from documents, and 43 cases where the project budget was obtained.

5. Findings

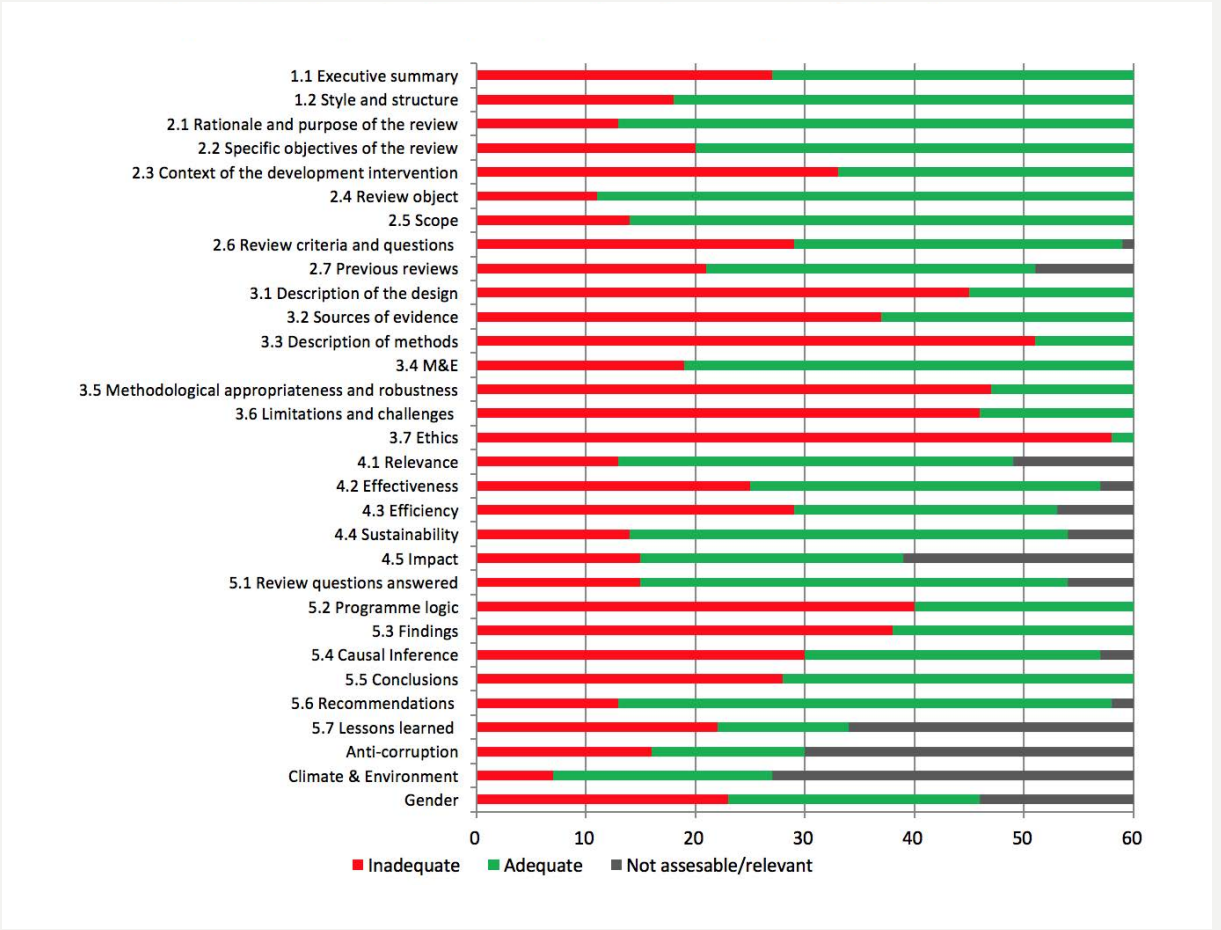
This Chapter presents the main findings of the evaluation, organised around the six evaluation questions. The findings draw on the data collected through the email and online surveys, the quality review and the five case studies.

5.1 WHAT ARE THE MAIN STRENGTHS AND WEAKNESSES OF REVIEWS OF NORWEGIAN DEVELOPMENT COOPERATION?

The main strengths and weaknesses against the 32 quality criteria defined in the review template and the 15 quality criteria defined in the TOR template are presented below, first for the reviews and then the TORs. Ratings were used to summarise the judgements made by our team. As explained in 3.1, each quality area was rated on a four-point scale with 1 representing poor quality, 2 inadequate quality, 3 adequate quality and 4 good quality.

The five quality areas of the review template can be summarised as follows: the first quality area covered the summary and the structure of the review, while the second covered review purpose, objectives, review object and review scope. The third area assessed methodology,

FIGURE 2: RATINGS FOR 32 QUALITY CRITERIA IN A SAMPLE OF 60 REVIEWS

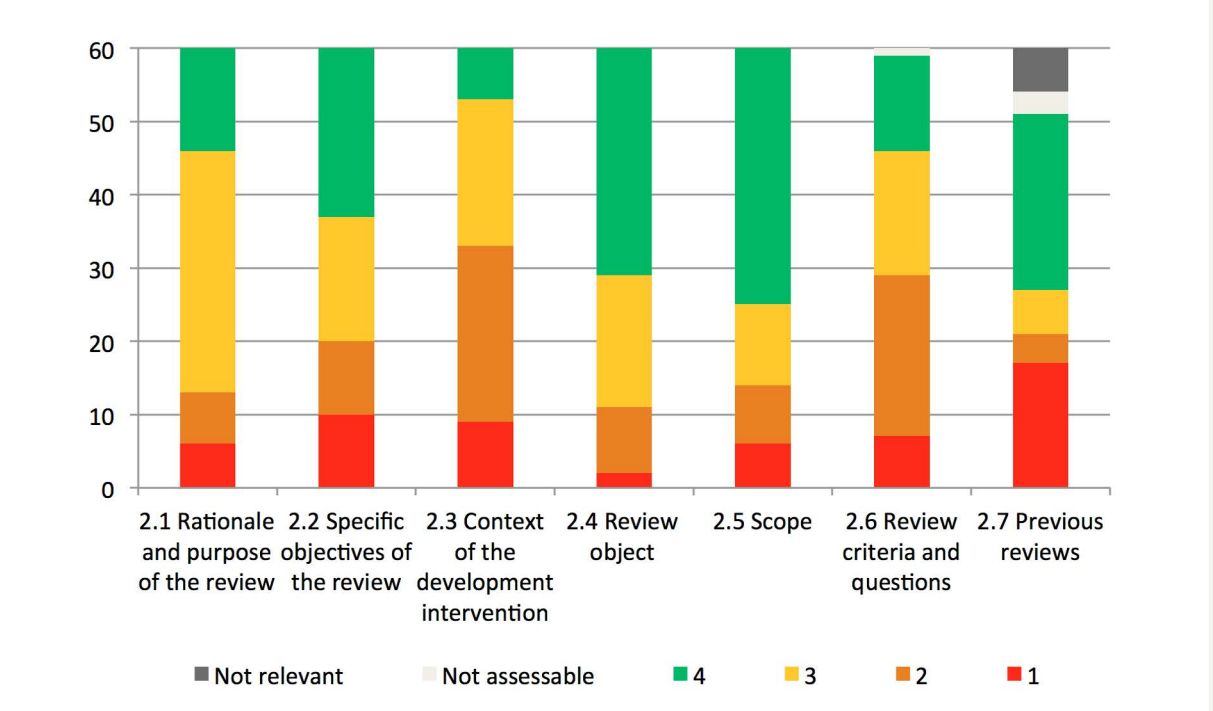


'Inadequate' is a combination of ratings 1 and 2, 'Adequate' is a combination of ratings 3 and 4.

with criteria rated for sources of evidence, description of methods, use of monitoring and evaluation (M&E) data, methodological robustness, limitations and ethics. The fourth area examined how well the OECD-DAC criteria were understood and applied (relevance, effectiveness, efficiency, sustainability and impact). The fifth area assessed the analysis findings, conclusions and recommendations, as well as the integration of cross-cutting themes.

The TOR template had three quality areas: the first covered nine quality criteria concerned with the review purpose, specific objectives, context, consideration of previous reviews, defining the object to be reviewed, scope, review criteria and questions, and finally feasibility. The second covered the review process, deliverables and quality assurance. The third section covered overarching and cross-cutting themes. The templates are in Annex 3, Appendix 1.

FIGURE 3: REVIEW RATINGS FOR SEVEN CRITERIA RELATED TO PURPOSE, CONTENT, SCOPE AND REVIEW CRITERIA



4: Good quality 3: Adequate quality 2: Less than adequate quality 1: Poor quality

5.1.1. Review strengths and weaknesses

Figure 2 illustrates the quality scores found across all the 32 quality criteria examined for reviews. The ratings have been merged into three categories for ease of presentation: inadequate (1 and 2), adequate (3 or 4) and not relevant/assessable.

Figure 3 illustrates the considerable variation found in the different quality areas. The highest quality scores were found in areas concerned with stating the purpose of the review, defining

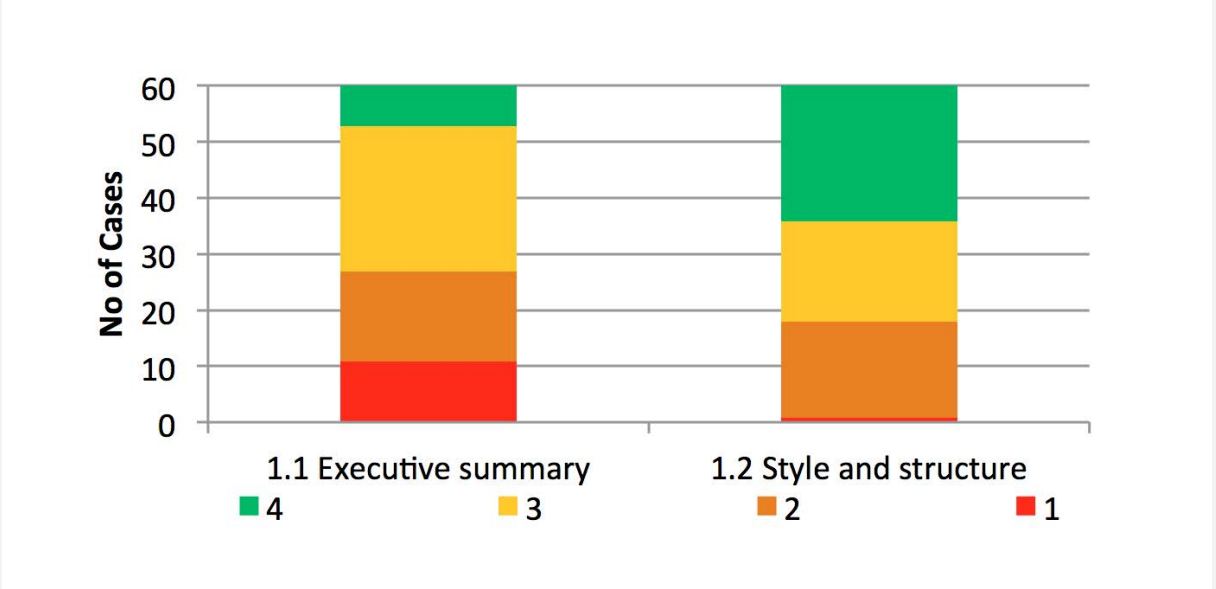
the object to be reviewed, answering the questions posed in the TOR, and making useful recommendations. Lowest quality scores were found in areas such as describing the methods to be used in the review, dealing with ethical issues and examining the programme's logic.

The following paragraphs present the detailed ratings for the sample of 60 reviews by six quality areas.

Quality area 1. Report summary and structure

Figure 4 shows that the balance between adequate and inadequate ratings for the ***executive summary*** was almost equal (32 were adequate and 28 inadequate). Of the 12 cases rated 1, this was because the reports did not include an executive summary. For those rated 2 and 3, the executive summary was partially complete. In these cases, the background (purpose, objectives of the review and context) was sometimes noted as missing, however the methodology was the most common piece of key information not included.

FIGURE 4: REVIEW RATINGS FOR EXECUTIVE SUMMARY AND REPORT STRUCTURE



4: Good quality 3: Adequate quality 2: Less than adequate quality 1: Poor quality

For ***style and structure***, 42 reviews were given an adequate or good (3 or 4) score, indicating that this was in general a strength of the reviews examined. Where they were rated as 1 or 2, this was mostly due to an illogical or confusing structure. The main example of this was a lack of clarity regarding the presentation of the findings, conclusions and recommendations. A couple of our reviewers noted that the report was poorly written, although this did not appear as a major issue across the sample.

Quality area 2. Review purpose, scope and questions

Rationale: 47 of the 60 reports were given a 3 or 4 score, indicating that explaining the rationale for the review was in general a strength (Figure 3). Those rated as a 1 were given this rating because of a lack of information across the quality sub-areas (why, when, for whom, how). Where there were weaknesses (for the reports rated a 2 or 3), in over half of cases it appears that this was because the 'for whom it is undertaken?' was only explained in a basic way, or there was insufficient information about how the report will be used.

40 reviews described the ***specific objectives*** reasonably well and 49 covered the ***object*** under study adequately or better. The attention to ***context*** was weaker: over half of the reviews were rated as either a 1 or 2. Weaknesses within the description of context related to each of the quality sub-areas assessed (policies, development context and cross-cutting themes). ***Scope*** was well addressed in 46 cases, usually drawing on the wording set out in the TOR.

Detailing the review ***criteria and questions*** was not well addressed in half the sample. For many, the OECD-DAC criteria and cross-cutting themes were either missing, incomplete or not made explicit, and the questions themselves were not clear or easily answerable. The best practice examples (see Annex 10) systematically and explicitly addressed the appropriate questions in the report.

Quality area 3. Methodology

The quality area covering ***methodology*** showed the lowest scores in general across the sample of 60 reviews (Figure 5). This is a critical aspect of an evaluation since without a sound methodology, it is very difficult for the user to have confidence in the reliability of the findings. The majority of reports had only a very limited discussion of methodology and some had none at all. There was little attempt in the majority of cases to explain the methodology used or indicate how it may or may not affect the validity of their findings.

45 reports (even those with fuller explanations of the methodology used) failed to conceptually embed their approach within a broader analytical or conceptual framework. Similarly the review design – the logic and structure of their approach – was often neglected. It was common for the reports to highlight the sources of evidence they relied upon, for example listing the documents they referred to and the people they interacted with. It was rare, however, for the sampling strategy to be well explained, and when sampling was referenced it was only to state in very basic terms who they included in their study, without the statistical implications or the question of bias being explored.

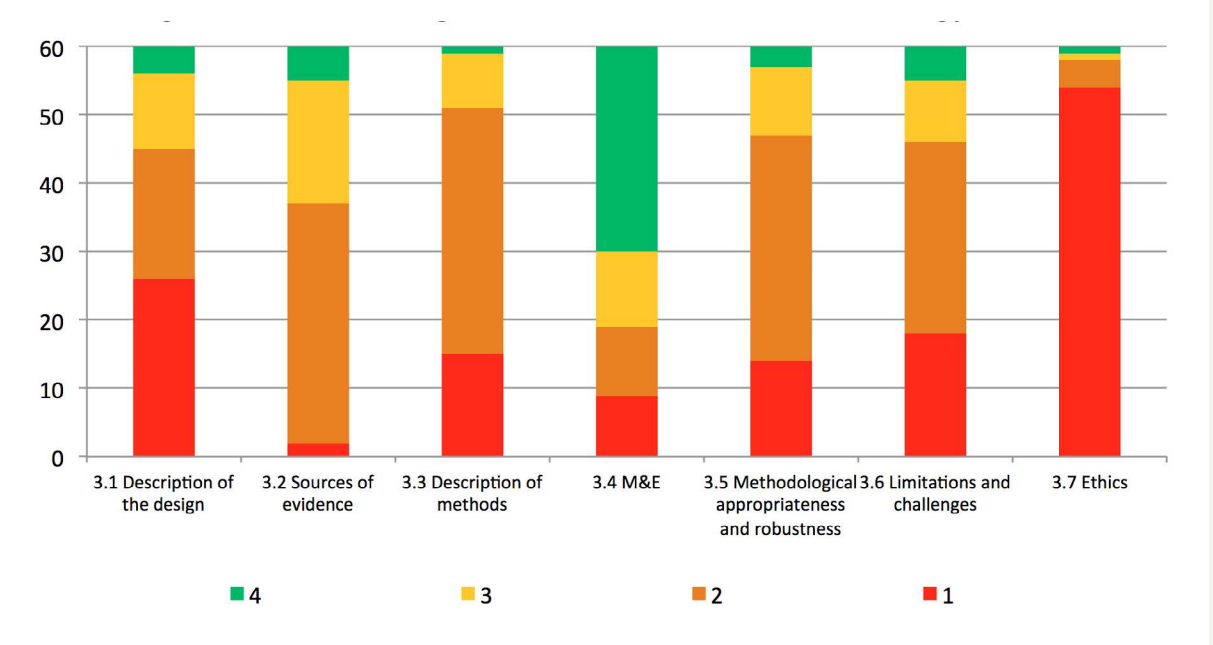
Although most reports included some description of the methods used, more detailed information concerning where and when these were used, and the rationale for so doing, was missing in 51 cases. Gender sensitive information was very rarely discussed in terms of methods, and more generally the mechanics of data analysis were also rarely examined.

Most reports appeared to choose methods that were broadly relevant for their purpose – that is standard tools such as key interviews, focus groups or even the triangulation of sources were frequently mentioned; however, they were inadequately explained. The link between the methods and evaluation questions was rarely explicitly established (with some notable exceptions). There was evidence of multiple lines of evidence being used with some triangulation; however, the logic of this was left unexplained. 41 reports did reference the project's M&E system, as well as critically discuss its quality and draw on the data.

45 reports were weak in presenting the ***limitations*** of their methods, which is important given the gaps outlined above. While some mentioned problems around missing or insufficient data, the implications of this for the findings were not then explored.

Finally, the quality area concerning ***ethics*** was the weakest of all the areas assessed. Only 2 of the 60 reports considered the issue of ethics adequately, and the vast majority were silent on questions such as confidentiality, respect for interviewees' rights and dignity and of giving feedback.

FIGURE 5: REVIEW RATINGS FOR CRITERIA RELATED TO METHODOLOGY



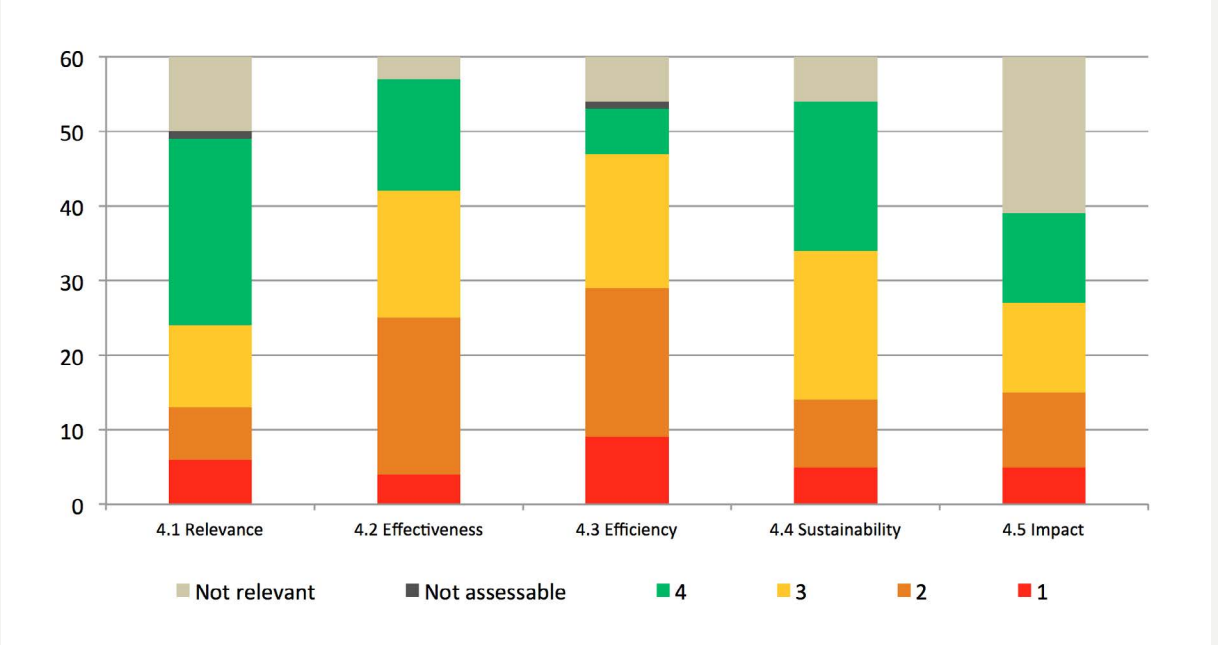
4: Good quality 3: Adequate quality 2: Less than adequate quality 1: Poor quality

Quality area 4. Understanding of OECD-DAC evaluation criteria

The assessment covered five common OECD-DAC evaluation criteria: relevance, effectiveness, efficiency, sustainability and impact (Figure 6). Where the TOR did not specifically ask for one of these DAC criteria to be examined by the review, the assessment marked this case as 'not relevant'. Where the TOR were not available, and the DAC criteria were not addressed in the review, it was marked as 'not assessable'.

Relevance was for the most part well understood by 36 of the 49 reviews rated (rated 3 or 4). Where it was rated poorly (1 or 2) the reasons given related to a generally poor understanding of the concept, having been requested in the TOR but not directly addressed in the report, or more likely it was only partially examined. This might mean it was poorly considered in relation to particular stakeholders, for example donors and end users; however, there was not one group in particular that appeared overwhelmingly neglected.

FIGURE 6: REVIEW RATINGS COVERING OECD-DAC EVALUATION CRITERIA



4: Good quality 3: Adequate quality 2: Less than adequate quality 1: Poor quality

Effectiveness was examined well (rated 3 or 4) by just under half the sample (25 out of 57 rated reviews). Where rated poorly (1 or 2), this was due to a variety of reasons including a poor understanding and application of the concept, the application of the concept not being systematically linked to outputs/outcomes, and a limited analysis or risk. Some reports confused outputs for outcomes. Others did not systematically link the indicators to the evidence presented.

Efficiency was rated less well than the other OECD-DAC criteria, with 29 out of 53 rated reviews given a 1 or a 2 rating. Where it was rated poorly, this was primarily due to a lack of consideration to alternative delivery modalities, as well as insufficient depth of analysis (for example the value for money/financial analysis being weak).

Sustainability was rated well across the sample with 40 out of 54 rated reviews given a 3 or 4. Where it was rated poorly (1 or 2) this was due to a general lack of depth to the analysis, a lack of consideration of environmental sustainability and, perhaps most fundamentally, a lack of direct judgement/analysis about the sustainability of the project.

Impact was considered in 39 cases, and was not judged as a relevant DAC criterion in the other 21. Impact was fairly well understood within this sub-sample, with 24 reviews rated 3 or 4. Where impact was rated 1 or 2 this was due to a variety of reasons including insufficient discussion and analysis, reliance on speculation, a conflation of outcomes and impact and, above all, a lack of consideration of end users.

Quality area 5. Quality of analysis, findings, conclusions and recommendations

This critical area examines if the evaluation questions are answered, how the findings are arrived at and how these are then linked to the conclusions, recommendations and any lessons. Overall there was a mixed level of performance across the quality criteria (Figure 7).

For the first criteria, whether the review ***answered the questions*** set in the TOR, there was reasonably good performance with 39 reviews out of 54 rated scoring 3 or 4. Some cases were not assessable because the TOR were not available to understand which questions were to be addressed. Others though did not fully address all the questions set.

Consideration of ***programme logic*** was a weakness, with 40 out of 60 reviews scoring 1 or 2. Although the programme design was discussed, the logic and assumptions underlying the design were not critically examined. There was also a lack of consideration of wider evidence and literature.

We rated over half of the reviews (38 out of 60) poorly (1 or 2) for the ***findings*** criteria. Weaknesses related to all of the aspects assessed in this quality criteria including the ability of the review to demonstrate a clear line of evidence, triangulation and gaps/limitations. Poorer reviews tended to have uncritical judgements and insufficient supporting evidence. Where we noted a weak line of evidence, a common explanation was the lack of sufficient discussion in the review about the data and methodology or insufficient triangulation. For many reports (even in those rated 3) there was only a limited discussion of gaps and limitations in the data and the significance of this.

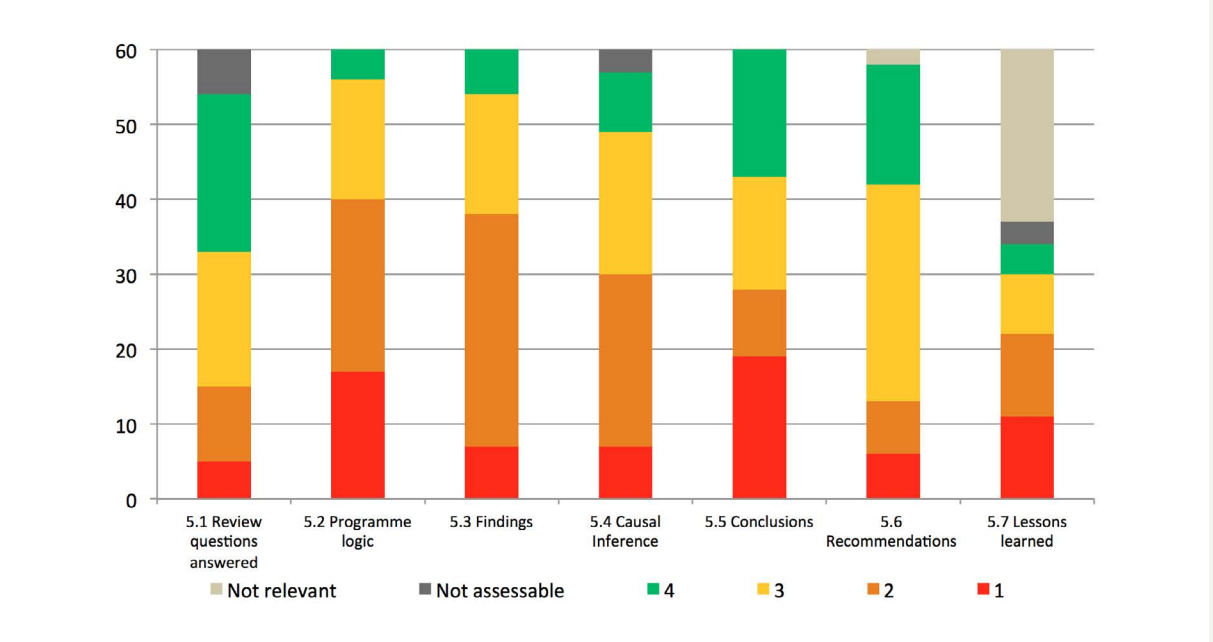
Causal inference was rated with mixed performance, with 30 out of 57 rated 1 or 2 and 27 as 3 or 4. The best examples clearly distinguished between outputs, outcomes and impacts. Where this was not done, it might have been due to the fact that in the majority

of reports no logical framework for the project was provided. Lower ratings also commonly occurred because of the lack of discussion on attribution and limited or no consideration of alternative factors causing results.

We rated just under half of the reviews as poor (either 1 or 2) for the ***conclusions***. The primary reason for a low rating was that no separate conclusions section was given. Sometimes conclusions were made in the findings section but without being adequately distinguished. Conclusions were sometimes also inadequately distinguished from the recommendations.

Of 58 the reviews rated for their ***recommendations***, we rated 35 as adequate (3) or good (4), indicating that this was in general a strength within the sample. Recommendations in these cases were judged to be relevant, actionable and targeted and to follow logically from the findings. Where there were weaknesses, however, these related to a number of areas including the recommendations being too brief, being disconnected from the findings, being disconnected from the conclusions and there being too many recommendations. A very common comment was that recommendations were not timed or prioritised.

FIGURE 7: REVIEW RATINGS FOR QUALITY OF ANALYSIS, FINDINGS, CONCLUSIONS AND RECOMMENDATIONS



4: Good quality 3: Adequate quality 2: Less than adequate quality 1: Poor quality

Nearly half of the sample of 60 was not assessed for ***lessons learned***. This was because either the TOR did not request lessons, or there was no section on lessons to assess. Of the remaining 34 cases,

we rated 22 as of inadequate or poor quality (1 or 2), because they were not separate but mixed in with conclusions and recommendations, or did not contribute to general learning beyond the project. Seven cases did not set out any lessons, even though lessons were requested in the TOR.

Examples of useful wider lessons are discussed separately in Annex 9.

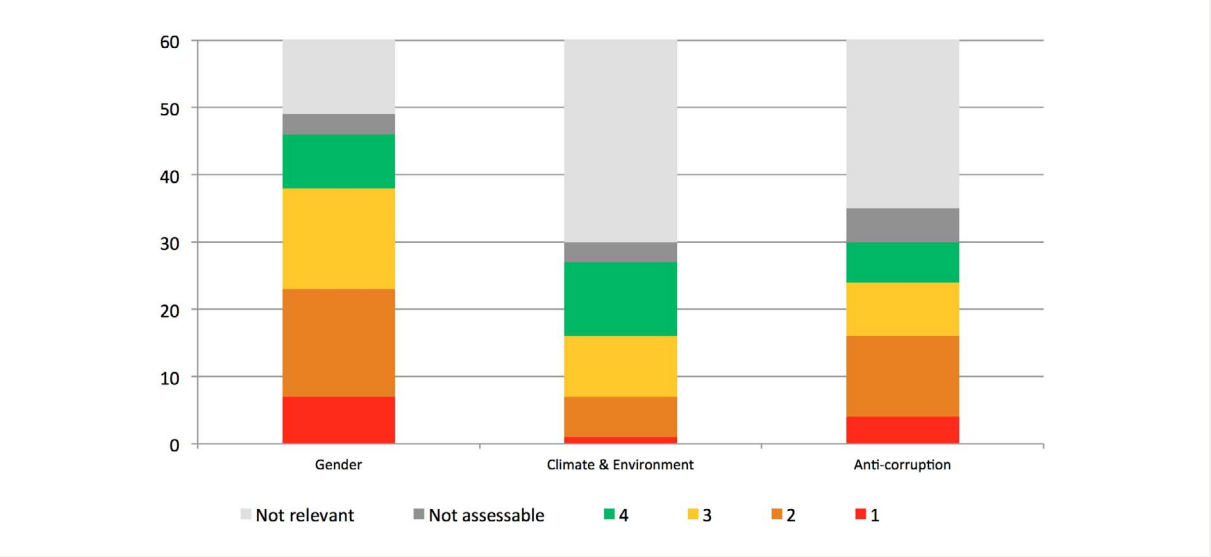
Quality area 6. Treatment of cross-cutting themes: gender, climate and anti-corruption

Where the TOR did not call for these cross-cutting themes to be addressed, the reports were not judged in terms of their quality but classified as 'not relevant'. In addition, where the TOR were not available, the reports were rated as 'not assessable' if they did not mention these themes.[26]

In the remaining 36 cases, ***gender*** as a cross-cutting theme was deemed weak in half of the reviews (Figure 8). Many of the reports that were rated 1 or 2 either had no or only partial consideration of gender. We noted that gender was, for example, included in the findings but not in the conclusions and/or recommendations. We also noted that there was a lack of depth to the analysis, even where it might be a focus of the project reviewed.

Where ***climate/environment*** was rated (27 cases), the majority of reviews (20) were scored as adequate or good. Where we highlighted this theme as a weakness, it was

FIGURE 8: REVIEW RATINGS FOR TREATMENT OF CROSS-CUTTING THEMES



4: Good quality 3: Adequate quality 2: Less than adequate quality 1: Poor quality

26 The TOR were expected to include these themes in all situations, and were rated accordingly. But the study only rated reviews where the TOR specified that these themes they should be addressed.

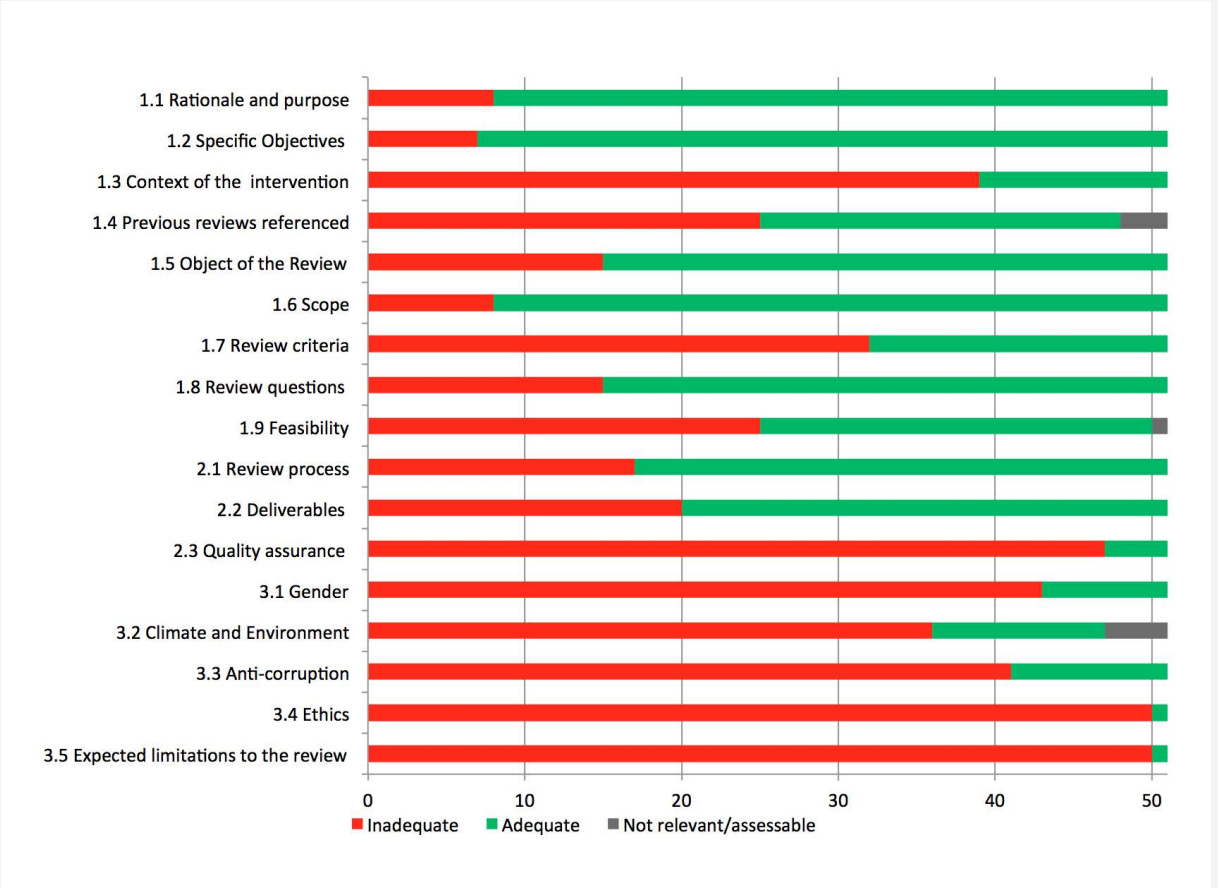
either because it was asked for in the TOR but was not then considered in the review, or it was inadequately considered (for example included in the findings but not in the conclusions).

Where ***anti-corruption*** was rated (30 cases), around half of reviews scored well (3 or 4). Where poorly rated, it was simply inadequately discussed, or in some instances although financial management was discussed, anti-corruption was not.

5.1.2 TOR strengths and weaknesses

Figure 9 illustrates the quality scores across the 15 criteria. Overall, we found the quality of how TORs addressed the review rationale, purpose and scope to be adequate or good, as was the quality of the description of the review process and deliverables. Areas with lower quality ratings included the context, review criteria, cross-cutting themes, ethics and limitations. These gaps fit with the weaknesses found in the reviews, especially in regard to context, themes, ethics and limitations. The TORs did not generally specify the methodology to be used. Given the major weaknesses found in the quality

FIGURE 9: RATINGS FOR 15 QUALITY CRITERIA IN A SAMPLE OF 51 TORS*



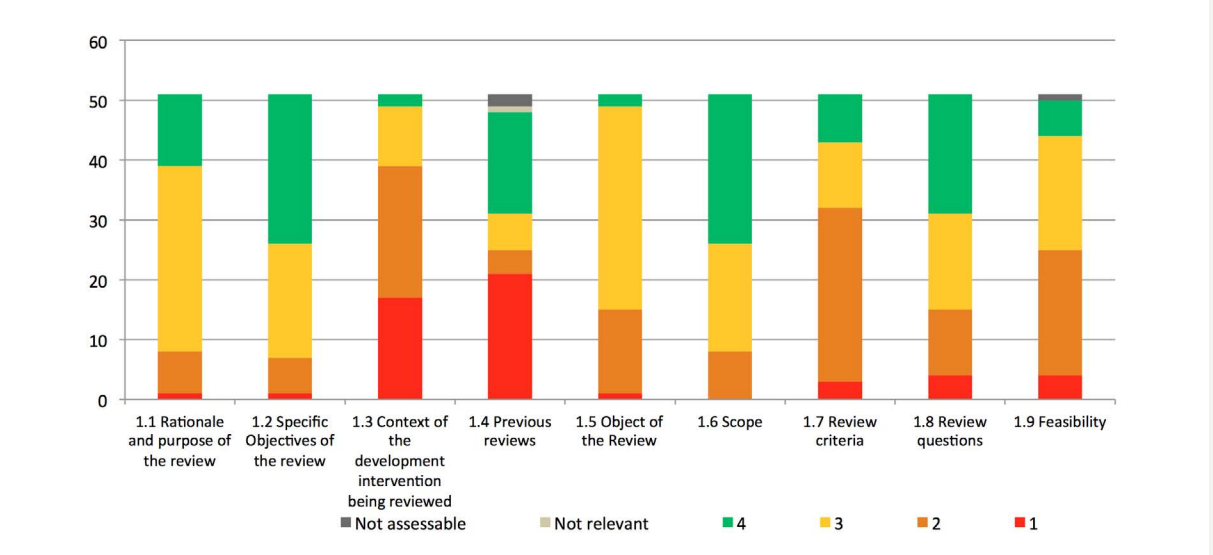
*'Inadequate' is a combination of ratings 1 and 2, 'Adequate' is a combination of ratings 3 and 4.

of review methodology, this may be an important gap in the TORs: if TORs had specified the approach to methodology in more detail then this could have prompted greater attention to this issue in the reviews themselves.

Quality area 1. Review purpose and objectives
Strongest quality criteria in the TORs were: ***defining the rationale and purpose***, and setting out the ***specific objectives*** of the review, ***describing the scope*** and setting out the questions to be answered (Figure 10). For the ***description of the review object*** (the nature of the project to be studied), details regarding the period, budget and geographical area were typically given, but it was very common for the intervention logic in particular to be unexplained. A description of the project ***outcomes*** was also missing. Though the main (organisational) ***stakeholders*** were identified, the ***organisational*** set-up was explained in only very basic terms.

On the other hand, often the ***context*** was poorly described, the definition of the appropriate ***OECD-DAC criteria*** was weak and the ***feasibility*** of doing the review (comparing the scope with resources) was poor. Where some context had been given, it was usually only related to the immediate organisational background relating to the programme or implementing organisation. The wider development context was neglected, as was the wider policy environment. Where we rated ***feasibility*** as poor, this was because of a lack of adequate resources in terms of time frame or days allocated, especially for fieldwork. ***Review criteria*** (whether the five OECD DAC terms: relevance, efficiency, effectiveness, impact and sustainability, or the three cross-cutting

FIGURE 10: TOR RATINGS CONCERNED WITH REVIEW PURPOSE AND OBJECTIVES



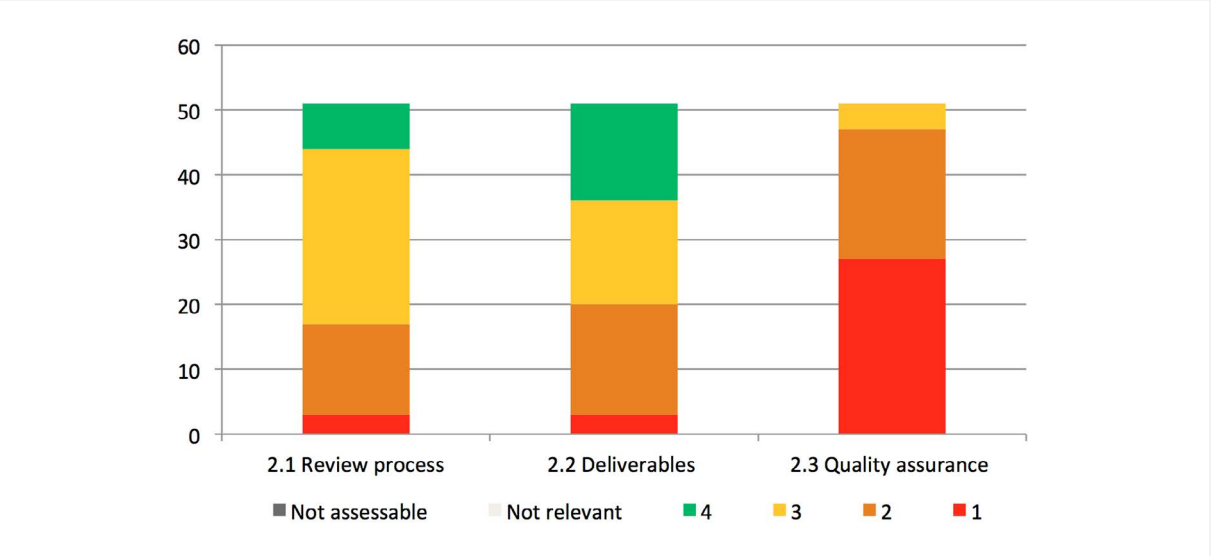
4: Good quality 3: Adequate quality 2: Less than adequate quality 1: Poor quality

themes: gender, anti-corruption or climate) were inadequately covered or missing in two-thirds of cases. Too high a number of ***questions*** was also a common weakness, though to a lesser extent. We noted that this was often hard to assess because the ***budget*** or specific day allocations were not made clear in the TOR.

Quality area 2. Review process

Figure 11 shows that TORs were mostly good at setting out the ***process*** of implementing the work and the ***deliverables***, but the majority neglected how the review should be ***quality assured***. Although some guidance was usually given around data collection and validation, there was a lack of clarification around having an inception stage. Furthermore, across the majority of reviews, roles and responsibilities were not fully explained.

FIGURE 11: TOR RATINGS FOR REVIEW PROCESS, DELIVERABLES AND QUALITY ASSURANCE

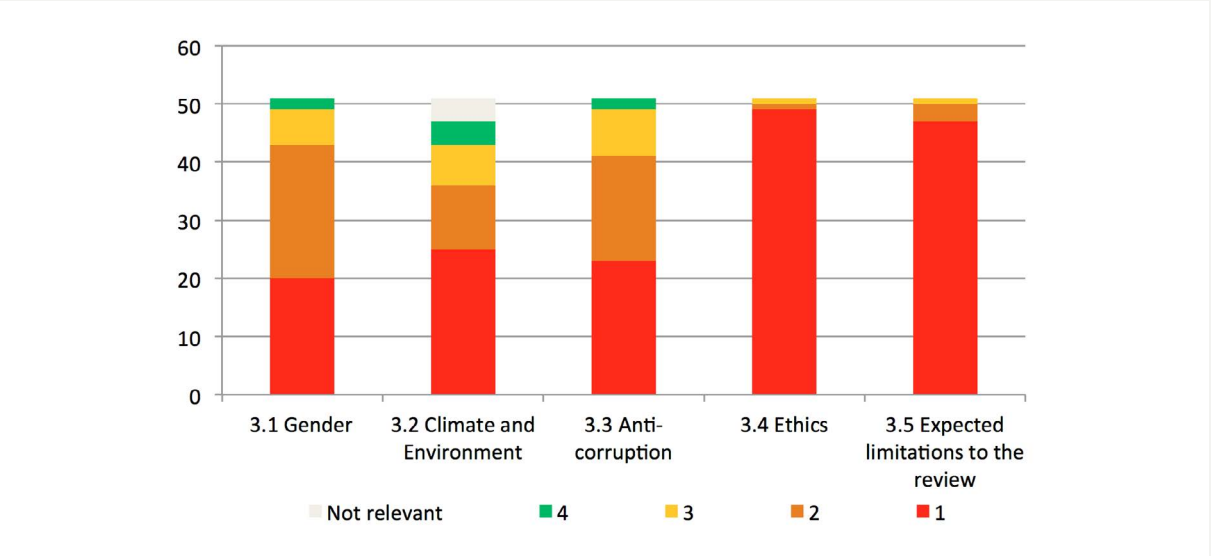


4: Good quality 3: Adequate quality 2: Less than adequate quality 1: Poor quality

Quality area 3. Overarching and cross-cutting themes

Generally the application of Norway's overarching and cross-cutting themes were the weakest aspects to be covered in the TORs (Figure 12). Most TORs did not mention gender, climate or anti-corruption, and almost none touched on ***ethics*** or defined the expected ***limitations*** that the review may face. In the majority of cases, ***gender*** was not well examined. Where it was mentioned, it appeared either in the introduction (regarding context for example) or given a brief mention in the review questions or focus, as if indicating the inclusion of gender as a cross-cutting theme – but not engaging with it in any substance. ***Climate*** was similarly neglected. Where it was included, it was most likely to be mentioned in the questions, but only in a cursory and non-substantive way. Similar to gender and the climate, ***anti-corruption*** was often neglected. Where it was included, it was described in terms of 'financial management', rather than in a broader sense.

FIGURE 12: TOR RATINGS FOR OVERARCHING AND CROSS-CUTTING THEMES



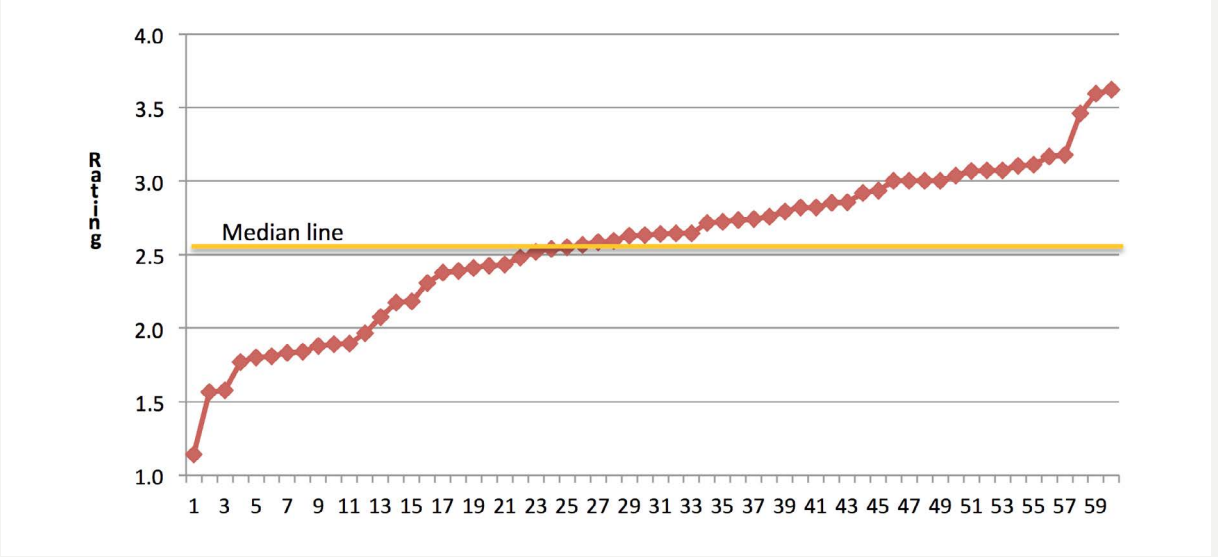
4: Good quality 3: Adequate quality 2: Less than adequate quality 1: Poor quality

5.1.3 Overall quality rating

For the purposes of examining what factors might influence review quality, an analysis was done to compare the review characteristics using an *overall quality score* that reflected the combined influence of the various quality areas assessed in the quality review.

Results for Reviews: For the 60 reviews, the average overall ratings ranged from 1.15 to 3.77. The majority (28 reviews) achieved an above mid-point average rating (with a rating over 2.5), while 22 were rated below this (Figure 13). Most reviews (45%) were rated in the 2.5–3.0 range but over a quarter were rated in the 3–3.5 range. For reviews, further analysis by quality area showed that the factors most closely associated with the overall score of a report were those in Section 5 of the review template devoted to analysis, findings, conclusions and recommendations.[27] This includes quality criteria concerned with causal inference, programme logic, the robustness of the findings and how well recommendations and lessons link back to these findings. As noted earlier, some aspects in this section were better addressed, such as whether the evaluation questions were correctly answered. But the most critical area related to review quality was how well the evidence gathered in the review was interpreted and analysed. This required sound causal links between the data and findings, and a clear progression from outputs to outcomes and impacts, and consideration of attribution. Where these steps were pursued more rigorously, the review would then be more likely to produce more reliable conclusions and recommendations. The quality of the areas concerned with cross-cutting themes was not so strongly associated with the other sections, and this indicated that the quality of these aspects varied independently of the other quality dimensions; that is, a high or low level of quality in the cross-cutting themes (related to gender, climate or anti-corruption) could occur irrespective of the results in the other quality areas.

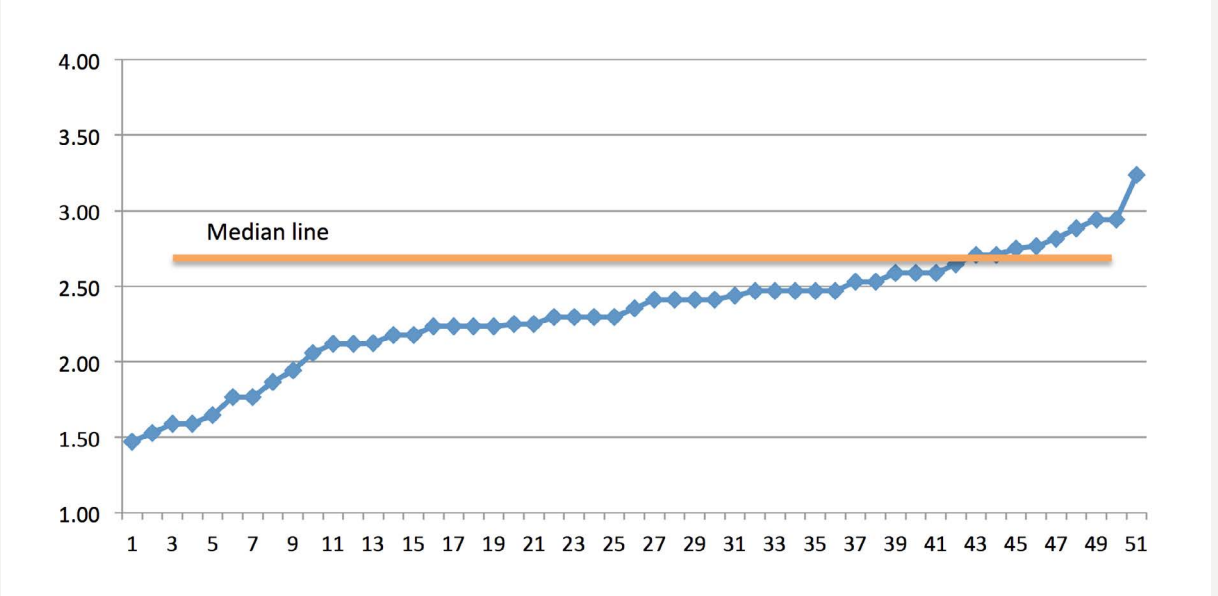
FIGURE 13: OVERALL RATING OF REVIEWS



27 The influence of difference sections was examined in a separate correlation analysis of ratings from the review exercise to see which quality sections were most aligned with the overall rating.

Results for TORs: For the 51 TORs in the sample, the overall average scores across 15 quality criteria ranged from 1.47 to 3.24 (with a score of 1 being interpreted as very poor quality and 4 as very good quality). The majority of TORs (35 out of 51 or 69%) fell below the mid-point of 2.5, implying that they were of poorer quality based on the review template (Figure 14). Only 31% achieved an average rating of 2.5 or higher, implying better quality.

FIGURE 14: OVERALL RATING OF TORS



5.2 WHAT FACTORS EXPLAIN VARIATIONS IN REVIEW QUALITY?

This section explores what independent factors collected in the quality review might influence quality.

5.2.1 Review and TOR quality by commissioner, region, type of review and sector

Overall ratings for TORs and reviews were analysed by commissioner, region, type of review and sector. The results were statistically tested to assess whether apparent differences in scores from the sample were likely to be true for the population as a whole in 2014, and given the similarity observed with the pool of reviews/evaluations over the period 2012–15 as shown in Annex 6, whether the sample results were likely to be true over the whole period. Such statistical testing was also important given the small sample size of 60 (reviews) and 51 (TORs), and the very low frequency counts for some categories.

No significant difference was found between the type of review commissioner and TOR or review rating. A test of variance showed that there was no statistically significant differ-

ence between the four types of commissioner in terms of the quality score of their TORs.[28] Reviews commissioned by MFA and partners also appeared to show higher quality average ratings. But again this was not found statistically significant (Annex 6, Tables 6).

No significant difference was found between the region where the review took place and either TOR or review quality score. In looking at TOR and review quality by region, there was some indication that reviews from Africa, Europe and Asia had a higher overall average quality score. However, the small sample sizes for most regions meant that any apparent differences were not statistically significant (Annex 6, Table 7).

No significant difference was found between the different types of review and either TOR or review quality score. In exploring the different types of reviews (Annex 6, Table 8), TORs showed some variation in quality scores, but there was no statistically significant difference between the four types. When end reviews ratings were compared with mid-term and review ratings, and the two evaluation ratings were excluded, then the end reviews *did have* a statistically significant higher quality score (F score of 3.94 against a minimum critical value of 3.2), implying that ***end review TORs might receive greater attention and care in their drafting.*** The quality areas where end reviews ToRs show higher quality were the review process, deliverables and quality assurance, and also scope and criteria.

No significant difference was found between different sectors and review quality scores. In terms of quality score differences by sector, scores appeared to show some variation between sector (based on 'Target area', a simplified categorisation provided by Norad for this exercise) (Annex 6, Table 9). But when tested for significance, there was a low probability that these differences were not due to chance.

5.2.2 Review and TOR quality by resources provided

There was evidence that those projects with larger budgets had higher quality scores for both TORs and reviews. A comparison of TOR and review quality ratings against the agreement budget showed a positive but very low degree of correlation between each of these two ratings and the budget allocated to the project. However, when ratings were analysed against the project budget as extracted from the review documents (TOR or review report), the relationship was stronger with a correlation of 0.44 for reviews and 0.24 for TORs. The difference arises from the fact that agreement budgets refer to the whole period of an intervention while the project budget reported in the review or TOR may in some cases refer just to the particular phase that was under review. Annex 6 Figure 2 illustrates this.[29]

28 TORs prepared by partners had a slightly higher overall average rating compared to those prepared by MFA, Embassies or Norad, but the sample was too small for this difference to be statistically significant.

29 The analysis using a 't' test of paired values gave a significant result, therefore suggesting that this relationship observed in the sample is likely to be true in the wider population and that projects that cost more are likely to have TORs and reviews of higher quality.

Larger projects tended to have reviews with bigger budgets. The study sought to compare project budgets against review budgets from the sample of 60 cases. Only 16 observations were available from documents where data for both these cost figures were found. Despite this small sample, there was *nevertheless* a significant and positive statistical relationship between the project budget and evaluation budget in the 2014 sample (Annex 6, Figure 3). The correlation coefficient observed was 0.41, and thus it seems that larger projects tend to commit more funds for reviews. Given the relatively small sample, however, it would be valuable to test this conclusion on a larger data set.

Reviews that had more days allocated for the work tended to score higher on quality. A comparison of the level of effort (in terms of the number of days provided for a review) and the overall quality rating found a significant and strong positive relationship (with a correlation coefficient of 0.52) (Annex 6, Figure 4).[30]

30 Although the sample was small (27 cases), a 't' test indicated that this relationship was likely to be reflected in the wider universe of reviews.

5.2.3 Comparison of overall quality ratings for TORs with quality ratings for reviews

Higher quality TORs were associated with higher quality reviews. A final dimension that was examined to understand overall quality was to see if higher scored TORs were associated with higher scored reviews. Analysis did in fact show a significant and positive statistical relationship between the overall average ratings of the TOR and the ratings of the reviews, with a correlation coefficient of 0.46 for the 51 cases available (Annex 6, Figure 5). The 't' statistic was highly significant indicating that this relationship is likely to occur in the wider population. This seems to imply that it is worth investing in preparing a high-quality TOR in order to improve the eventual quality of the review.

5.2.4 Are reviews based on good data, methods and analyses likely to produce credible information about the programmes and their outcomes?

The quality review found that a significant proportion (over 65%) of the reviews conducted on Norwegian aid programmes were highly unlikely to contain sound methodological underpinnings that would support or produce credible findings. Often reports had only a very limited discussion of methodology and some had none at all.

Such details are often included in inception reports, since this is usually where in the review process that the approach to gathering evidence would be set out. However, in the 2014 sample of reviews, few appeared to have inception reports (only three were located following document search and the email survey). The issue was further investigated to see if inception reports (1) were asked for in the relevant TORs, or (2) if they were referred to in the review reports themselves. Subsequent analysis showed that only 23 of the 51 TORs mentioned the need for an inception report (and in some cases this was only a fieldwork plan).[31] In addition, of the 60 review

31 The TOR quality review template recorded whether an inception report was stipulated (quality area 2.1). From analysis of the review team's comments, only 23 out of the 51 TORs included the requirement that an inception report be prepared.

reports examined, just 16 referred to an inception report or phase. Furthermore, in the five case studies examined in detail, only one of them had an inception report.

The conclusion is that even if inception reports had been reviewed, the fact that the majority of reviews did not have them indicates that their influence on this study's judgement about methodology would unlikely be any different. In our view, the main report should still have sufficient information on the review methodology, *especially* if there was no inception report, to enable the reader to judge the quality of evidence presented.

The *quality area* covering methodology was judged to have the lowest level of quality across the sampled reviews, with an average rating across the six quality criteria concerned with methodology of 2.1. Further on the *analysis* side of the reports, while two-thirds of reports *answered the questions* set in the TOR, there was insufficient analysis of the *programme logic*. Although the programme design is often discussed, the logic and assumptions underlying the design were not critically examined and there was often a lack of consideration of wider evidence and literature.

As a consequence of the challenges faced in methodology and analysis,[32] we found that the quality of the *findings* presented in the reviews was poor, with 37 reviews rated as a 1 or 2. Weaknesses related to all of the sub-areas in this quality area (line of evidence, triangulation and gaps/limitations):

- A weak line of evidence was often pointed to, reasons for which include uncritical judgements and a lack of supporting evidence, or lack of sufficient explanation about the data and methodology or insufficient *triangulation.*

- For many reports (even in those rated as adequate), there was only limited discussion of *gaps and limitations* in the data and the significance of this.

- The quality area, *causal inference*, was rated with mixed performance, with 32 reviews rated 1 or 2 and 28 as 3 or 4. The best examples showed the likely linkages between outputs, outcomes and impacts.

- Lower ratings also commonly occur because of the lack of discussion on *attribution* and limited or no consideration of alternative factors causing results.

- It was noted that in the majority of reports no *logical framework* for the project had been provided. However, the construction of logframe, theory of change or similar is an important tool for establishing likely causal effects and in good reviews, this was developed even if the project did not have such a framework.

32 There was close alignment between reports that score poorly for methodology and for findings.

5.3 TO WHAT EXTENT DO REPORTS PRESENT ANY GENERAL LESSONS LEARNED WITH RELEVANCE BEYOND THE INTERVENTION UNDER REVIEW?

Given the programme-level focus and the relatively light-touch nature of many of the reviews in the sample, it is not surprising that ***the evaluation team considered there were few reviews providing significant new learning.*** The study identified just seven reviews that had lessons of this nature. These are set out in Annex 9 and briefly summarised here.

The lessons have been categorised as relating to ***programme design and delivery*** or to ***specific themes.***

Programme design and delivery:

- The value and pitfalls of partnership in design and delivery: Lessons regarding partnerships which the evaluation team deemed important and of wider applicability emerged from four reviews.[33] These lessons were in the areas of the significance of partnership for sustainability, the opportunities of high-level government involvement, and the value of partnership with complementary organisations. Alongside this, the negative consequences of weak partnerships were also highlighted.

- The importance of management as well as models: Two reports[34] highlighted how the quality of programme management was absolutely critical to programme success. Even when a project was well designed, it was the way in which it was managed that determined the success of the outcomes.

- Relationships between delivery organisations and the Norwegian aid administration can be both constructive and problematic: One review[35] offered a number of insights into the relationship between the Norwegian aid administration and the organisations that they fund. These reflected both constructive and problematic dynamics in the relationship.

Themes:

- The significance of media: Two reviews[36] highlighted the importance of the media in, first, gaining respect and recognition for the programme's work beyond the immediate organisations affected; and second, in widening ownership among project beneficiaries.

- Peace building/political transitions: The lessons from three reviews[37] focus on a number of areas, including the potential for future lessons learned in Palestine,

33 REDD+Initiatives in Costa Rica (review 48), Lake Chilwa Basin Climate Change Adaptation Programme (review 124), Norwegian democracy support via political parties programme (review 75), Expanded Programme on Immunisation (EPI) in the Zambézia province, Mozambique (review 147).

34 REDD+Initiatives in Costa Rica (review 48), Mid-Term Review for Lake Chilwa Basin Climate Change Adaptation Programme (review 124).

35 Norwegian democracy support via political parties (review 75).

36 Benguela Current Commission (BCC) Science Programme (review 16), Lake Chilwa Basin Climate Change Adaptation Programme (review 124).

37 Palestinian Negotiations Support Project (review 199), Monitoring Nepal's Peace Process and Constitution Drafting programme (review 157), Norwegian democracy support via political parties (review 75).

the value of political monitoring, difficulties working in post-Soviet countries, and measuring outcomes in such contexts.

5.4 FROM THE PERSPECTIVE OF STAKEHOLDERS, TO WHAT EXTENT ARE REVIEWS TIMELY, AND PRESENT RELEVANT AND REALISTIC RECOMMENDATIONS?

In this section, we examine what factors are considered in assessing if a review is timely or not and the extent to which stakeholders perceived reviews to be timely. We then consider the extent to which stakeholders perceived that the recommendations from the reviews were relevant and realistic and how this influenced their use. The section draws on data collected from the surveys (online and email) and the key informant interviews for the five review case studies.[38]

5.4.1 Timeliness

The extent to which a review is considered to be timely was linked to factors which influence the use of the review. Reviews were considered timely when the timing of conducting the review contributed to its use. There were several factors mentioned of how timeliness influences use including:

- Extent to which the review was timed to be able to influence decisions on future funding (review 122), the next phase of a project (review 155) or policy decisions (review 155).
- Extent to which the review team was able to access information needed to conduct the review (review 184).

It was reported by members of the Norwegian aid administration that use was not confined to just them but was also considered in terms of other stakeholders such as government, policy makers or implementing partners. For example, the Nepal review (155) was perceived to be timely, in part, because the timing aligned with when the government was reviewing its energy policy and the government was thus able to use the review. The Pakistan review (184) was potentially timely for co-funders and implementing agencies, who were the intended users of the review, rather than the Norwegian aid administration.

Reviews were generally considered to be timely: although the sample size was small, 71% of respondents (n=24) to the online survey reported that reviews were timely in relation to their intended use[39] and three out of the five case study reviews (reviews 16, 155, 122) were considered to have been timely. For example, the project review of ProVert Integrated Green Education Programme in Madagascar (review 122) was considered favourably because it was conducted when decisions were being made about the next funding period. The Mid-Term Review of the National Rural and Renewable Energy Programme in Nepal (review 155) was used to influence changes in an ongoing programme and it was conducted at the

38 References to review numbers relate to their code number assigned in the Mapping Study. The numbers are listed in Annex 5.

39 Question 4.1 – To what extent are mid-term, end-term review or evaluations timely in relation to their intended use? On a scale of 1 to 4, with 1 being 'no, not at all' and 4 being 'Yes, very much so', 71% of respondents responded with a rating of 3. It is worth noting that no one gave a rating of 'Yes, very much so'.

same time that the government was revising its energy policy and, hence, was used by the government in the revisions.

However, two of the case study reviews (review 184 and 244) were not considered to be timely. In one case, this was because it was conducted after the decision had already been made to discontinue funding.[40] In the other case, this was because the review was carried out after the project had finished which made it difficult for the review team to access sufficient information and key stakeholders.[41]

5.4.2 Relevant and realistic recommendations

The quality of the recommendations was a key factor in use of reviews: 95% of online survey respondents reported that the ability of reviews to deliver concrete recommendations was a key factor in use of reviews.[42] Even though the survey response rate was quite low, the large percentage expressing this opinion highlights the importance of the recommendations being relevant to the intervention being reviewed and feasible to apply.

Responses from both the online survey and the case studies indicated that overall the reviews presented relevant and realistic recommendations: 67% of respondents (n=24) to the online survey indicated that the review recommendations were relevant and realistic.[43] Additionally, the recommendations were considered to be relevant and realistic in four (reviews 16, 155, 122, 184) out of the five case studies.

Recommendations were considered relevant and realistic by stakeholders because they fed directly into the needs of the users, both for the Norwegian aid administration and the grant recipient and they were able to be actioned to make positive changes to the programme and to government policy (review 122, 155).[44]

5.5 HAVE REVIEW FINDINGS, CONCLUSIONS AND RECOMMENDATIONS BEEN USED BY THE UNIT RESPONSIBLE FOR MANAGING THE GRANT?

Based on the conceptual framework set out in 3.2 to categorise the different types of use, this Chapter presents the main findings on the use of reviews by the unit responsible for managing the grant in the Norwegian aid administration.

40 244 – In the case of the End Review: Assistance in Management of Petroleum Resources in Timor (review 244), the decision had already been made to discontinue funding when the review was conducted. This timing influenced the design of the review – it was limited to a desk study which in turn likely contributed to low buy-in from grant recipients and low use.

41 184 – The End of Project Evaluation: Norway Pakistan Partnership Initiative (review 184) was carried out after the project had finished. This made it difficult for the review team to access sufficient, relevant information and stakeholders, and likely contributed to a low-quality review which had low use.

42 Question 4.10: Survey respondents' assessment of what factors are important in affecting the level of use of a review; 95% responded that 'Delivers concert recommendations for improving the project' is important/ very important.

43 Question 4.2: *To what extent are the reviews and evaluation recommendations realistic and relevant?* Respondents ranked reviews on a scale of 1 to 4 with 1 being 'No, not at all' and 4 being 'Yes, very much so'.

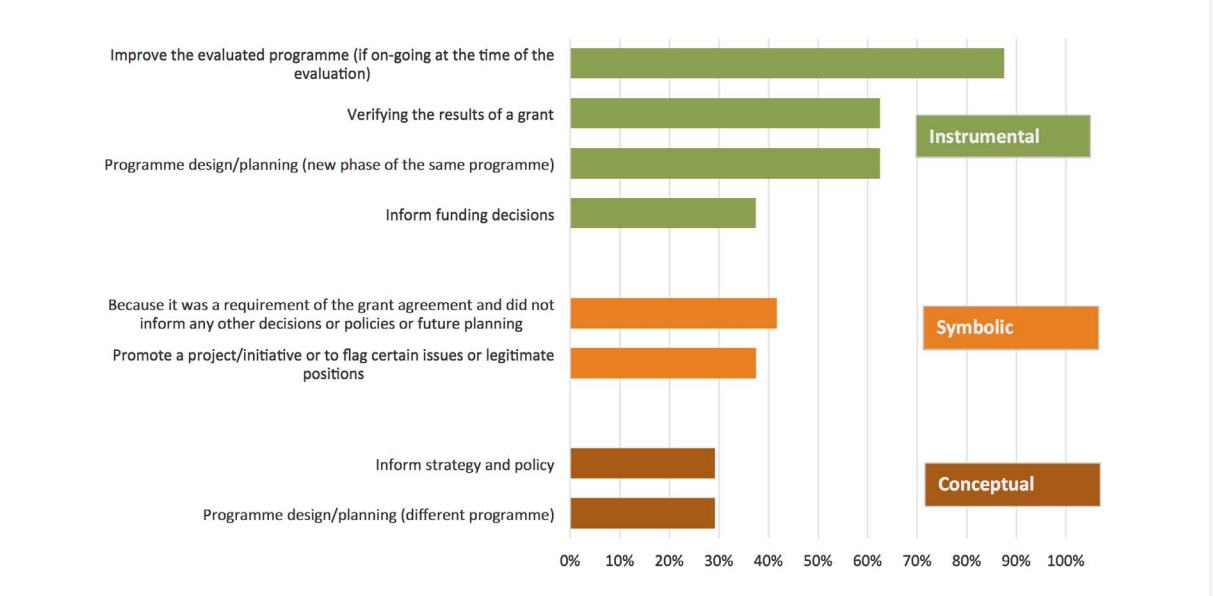
44 There was one case study review, End Review: Assistance in Management of Petroleum Resources in East Timor (review 244), where the recommendations were not considered to be relevant or realistic. This was because the recommendations did not respond to the TOR. The TOR explicitly stated that the review would not consider a possible extension of the programme. Nevertheless, the recommendations focused solely on supporting a project extension. Given the explicit request in the TOR to the contrary, the recommendations were neither relevant nor realistic since the Norwegian aid administration had already decided not to fund a further phase. The lack of relevant and realistic recommendations for the Norwegian aid administration may have also contributed to the low use of the review.

5.5.1 Findings

The online survey asked how respondents used reviews in their work.[45] Respondents were able to choose from eight options which we have categorised as instrumental, symbolic or conceptual use (Figure 15).

There was a high level of use of reviews by the unit responsible for managing the grant for these interventions. A significant majority of respondents to the online survey (91%, n=24) indicated that the reviews were useful or very useful.[46] When asked to provide examples of particularly useful evaluations, respondents cited 11 reviews which were used in a variety of ways including to make decisions about funding decisions and programme design and planning, and to prepare for field visits.[47] Some 76% of respondents to the email survey (n=34) indicated that reviews had been used. However, it is possible that this response under-represents use due to a low response rate to the survey stemming from rotation of staff in the Norwegian aid administration. Some of those who were contacted for the email survey indicated that they were not the right person for answering the questions posed.

FIGURE 15: CATEGORISATION AND USE OF REVIEWS



Reviews were most often used in instrumental ways. Reviews that can be used in instrumental ways play an important role in supporting grant managers in their day-to-day work. Given that grant managers are balancing large portfolios and have limited opportunities to visit the interventions they are managing, reviews that can provide them with the information needed to make important management decisions add value.

45 Online survey Question 3.1.

46 Online survey Q 3.2.

47 Online survey Q 3.4.

The three most frequently reported ways in which results were used were categorised as instrumental: to improve the programme being evaluated; to verify the results of the grant; or to influence the programme design (see Figure 15). Additionally, three out of the five case studies (cases 16, 122, 155) were used in instrumental ways.

For example, the Benguela end review (review 16) was used to inform a new intended cooperation phase with the partner including a detailed follow-up with the grant recipient based on the review recommendations. The grant manager subsequently found that the project had changed in accordance with some of the key recommendations. The Madagascar review (review 122) generated nine concrete recommendations, and these were actively followed up by Norad and the embassy. This resulted in comprehensive changes to the programme in its new phase and a substantial cut in the budget for the programme.

The second most common way in which reviews were used was to verify results of the grant. We found that 63% of respondents to the online survey used reviews to verify the results of a grant (Annex 8, Table 1). However, it is worth recalling here that almost 50% of 60 reviews in the quality review sample were rated as poor or inadequate in quality in assessing the *effectiveness* of the intervention, while 55% assessed *efficiency* to this standard, while the *findings* of 63% of the reviews were rated poor or inadequate. Thus, two out of three grant managers are verifying grant results using reviews, but a large proportion of these reports are rated as below standard in quality areas related to results.

There was evidence of symbolic use[48]. As seen in Annex 8, Table 1, 42% of respondents to the online survey reported that reviews were conducted because it was a requirement of the grant and 38% reported that the review was done to validate a position or decision that had already been made. Additionally, two of the case study reviews were considered to be of symbolic use (cases 184, 244). In the cases of the case studies, the symbolic use of the reviews was also linked to low use.

Conceptual use of the reviews was limited. Conceptual use was least frequently cited in the online survey and was only evidenced in two of the case studies (cases 16, 155). Use is considered conceptual when reviews are used to inform strategy, policy decisions and programme design that go beyond the intervention being reviewed. The low level of conceptual use of reviews indicates that use of reviews was largely confined to the intervention being reviewed.

The limited conceptual use of reviews could be due to the lack of lessons learned resulting from reviews and/or weaknesses in the Norwegian aid administration's system to catalogue reviews. Nearly 50% of the sample of 60 reports in the quality review were not assessed for lessons learned because either the TOR did not request lessons, or there was

48 Where an evaluation fulfils a bureaucratic or programming requirement rather having its own intrinsic merit.

no section on lessons to assess. Of the remaining 34 cases, 22 or 65% were rated as of poor quality because they were not separate but mixed in with conclusions and recommendations, or did not contribute to general learning beyond the project. Seven cases did not set out any lessons, even though lessons were requested in the TOR. This shows that a limited number of reviews include lessons learned and where lessons are included, they are generally of poor quality.

Additionally, 50% of survey respondents reported that it was difficult or very difficult to find and access reviews in the current management information system.[49] This can be triangulated with the evaluation team's own work to find reviews in the system. It was time consuming and difficult to find the reviews and related documents when accessing the archives, even when all the project and review information (file numbers, titles etc.) was known.[50] This makes it difficult for reviews to be accessed beyond those directly involved in the reviews and, thus, limits conceptual use of reviews.

However, it is worth considering whether reviews should be expected to be used in a conceptual way. Given the relatively limited budget and resources available for the reviews, it may be sufficient that they are used in instrumental ways, and it may be too ambitious to expect them to provide lessons that can be used to influence other initiative across the Norwegian aid administration or by other parties.

Two review case studies that were used in conceptual ways were also used in instrumental ways. This indicates that these reviews had added value both in improving the ongoing programme or policy, and in influencing broader thinking and programming beyond the specific intervention being reviewed. For example, the Benguela review also functioned as a source for learning in the Norwegian aid administration. It was used by the embassy to understand the impact of the Benguela Current Commission (BCC) and how Norway's support had assisted, and the information was fed to MFA and Norad. More widely, the review also fed into the 'Fish for Development' programme begun by Norway, and it has enhanced their competence in sustainability work on oceans. In 2015, a team from Norad came to the Norwegian Embassy in South Africa to look at the 'blue economy' and they used this review. The review contributed to the understanding of these issues regionally and internationally and to related political issues, e.g. the 2016 ocean conference in Washington, 'security at seas'. The political relations between the three countries was sensitive and the review addressed not just the scientific issues but the gains made in terms of tri-partite aspects of building consensus.

49 Online survey question 3.6.

50 The reviews can be filed as a case separate from the project itself. If someone is interested in collecting the evidence from reviews, but does not know that a review actually has been conducted, it would be very difficult to find in the system.

5.6 WHAT ARE THE MAIN FACTORS CONTRIBUTING TO QUALITY AND USE OF REVIEWS?

The following section looks at the factors which most impact the quality of reviews and evaluations and the factors arising from this evaluation that appeared to contribute most to the use of reviews.

5.6.1 Factors contributing to quality

While there were a wide number of issues that shape quality, those outlined below were the three for which there was the strongest evidence base.[51] In summary, these were:

- The quality of the review terms of reference;
- The level of resources (budget and level of effort) allocated to the review;
- The calibre of the review team.

The quality of the review TOR was an important factor in shaping the quality of the final report. As part of the analysis of the 60 quality reviews, the evaluation looked into the relationship between different factors and quality. A significant and positive statistical relationship was found between the TOR and review ratings (Figure 20).[52] Higher quality reviews tended to have higher quality TORs. This finding also came out strongly in the case studies. A good quality TOR was identified as essential to quality in all five case studies. Most review team leaders (cases 16, 122, 155, 184) and all five grant managers stated that a high quality TOR is a key factor for getting a high quality review. There was a common view of what a high quality TOR should contain across the case studies: a specific purpose and clear specifications on the required knowledge and experience of the review team. This finding resonates with another study of evaluation quality which have also found a relationship between TOR and review/evaluation quality.[53]

A key enabler for developing high quality TORs was having adequate time. A number of the case studies pointed towards a good quality TOR being one that has undergone extensive consultation with the relevant stakeholders. This was important to ensure that the right issues were being looked at, as well as fostering ownership and buy-in to the review process (see below for a discussion of the link between consultation and use).

The level of resources allocated to a review was an important determinant of quality. Based on the analysis of the 60 quality reviews, a significant and positive relationship was found between the budget and level of effort (number of person days allocated to a review) and the overall quality of the review report. Reviews with more resources were associated with higher quality. This finding was also echoed through the online survey, where 83% of respondents (20 out of 24) indicated that the level of resources for a review influenced its quality. This is consistent with

51 It should be noted that because of the small sample, more advanced forms of statistical analysis (such as regression or principle components analysis) were not deemed appropriate.

52 A scatter graph and regression shows a positive relationship, with a Pearson correlation of 0.46.

53 Global Affairs Canada (2016) Meta-evaluation of Global Affairs Canada's Decentralised Evaluations: FY 2009/10–13/14, August 2016, Ottawa.

a number of other studies on evaluation quality which have also found that resources and quality are linked.[54]

The calibre of the review team was a key determinant of review quality. This finding was supported across both the online survey and case studies. The online survey identified the quality of the team as the top factor in determining quality.[55] The case studies painted a similar picture. In all five, a well-qualified review team was associated with a high quality review (reviews 184, 122, 155, 16, 244).

The attributes of a high quality team were varied. Appropriate evaluation expertise was key.[56] Knowledge of the context, subject and project were also identified as important. In the Pakistan case study, for example, both the review team leader and the grant manager commented that a strong team with strong subject specialist and local knowledge and knowledge of the Pakistan government agencies who implemented the projects were key factors in explaining the quality of the review (Annex 7). Other slightly less specific, but no less important, attributes of a high quality review team were being adaptable, objective and professional.[57]

Interestingly, while other studies of evaluation quality have both an association between the calibre of the team and quality, as well as the size of the team and quality, no such relationship was found in this evaluation. Statistical tests on the data from the 60 quality reviews found no association.

There was limited evidence to suggest that stakeholders within the Norwegian aid administration believed that a robust methodology was important to quality. In only one case study was it mentioned that a comprehensive methodology was important for generating a high quality review (184).[58] Moreover, none of the online survey respondents identified methodology as a key factor in supporting quality.[59]

What emerged from the available evidence, particularly the case studies, was an understanding of quality within the aid administration weighted heavily towards usability and limited consideration of methodological rigour. High quality reviews are ones that have actionable insights in the findings, conclusions and recommendations. Across all of the case studies, reviews were considered high quality if they revealed something important and useful that could be implemented to improve the project.

54 Lloyd, R. and Schatz, F. (2015) op. cit.; Australian Department of Foreign Affairs and Trade (2014) op. cit.

55 14 out of 17 respondents that answered to the question 'overall, what are the most important factors that support good quality in reviews?' highlighted the quality of the review team and consultants as key.

56 A USAID study found that USAID evaluations with an evaluation specialist as part of the team were statistically of significantly higher quality (USAID (2013) Meta-evaluation of Quality and Coverage of USAID Evaluations 2009–12, prepared by Management Systems International).

57 Online survey.

58 Methodology was also mentioned as one of several key factors contributing to review quality by the grant manager in case 16.

59 In response to an open-ended question 4.11 from the online survey, 'Overall, what are the most important factors that support the eventual use of reviews?' No respondents identified methodology. The most common factor identified as reducing the quality of reviews was the weakness of the review team. This was described in terms of being incompetent, not knowing the context, not knowing the language, and 9 of the 15 respondents highlighted this issue. See Annex 8 question 4.11 for more detail.

A challenge with this is that grant managers and project implementers were sometimes making actions based on evidence that this evaluation assessed to be low quality evidence. So while there was an inclination to view reports as high quality if they offer practical recommendations, this could be divorced from an assessment of the underlying methodological rigour of the data collection and analysis. This picture was complicated further by the fact that 74% of respondents to the online survey felt that the methodology used in reviews was appropriate for the scope and objectives set out in the TORs. This suggested that within the aid administration there might actually be a sense that the methodologies being used were robust and therefore the evidence sound.

5.6.2 Factors contributing to use

While there were a wide number of issues which shape use, those outlined below are the four issues for which there was the strongest evidence base. In summary, these were:

- The quality of the review terms of reference and the delivery against these;
- The inclusion of realistic and actionable recommendations;
- The engagement of key stakeholders in the review process;
- The timing of the review.

A key factor contributing to the use of reviews was the formulation of high quality TORs and the delivery against this. All 24 respondents to the online survey believed that delivering on the specifications of the TOR was an important factor in explaining whether a review is used. This was also identified in three of the five case studies (reviews 122, 16, 155).
The two ways in which the quality of the TOR for a review links with the eventual use of the findings, conclusions and recommendations was through: stakeholder engagement and clarity of purpose. In two case studies respondents argued that engaging with key stakeholders around the development of the TOR helped generate ownership and buy into the review and that this laid the foundations for eventual use (reviews 122, 155). If stakeholders feel that a review was looking at the questions they were most interested in and that were most relevant to the decisions they needed to make, uptake was more likely. Similarly, in two other case studies, respondents argued that a focused TOR with a clear purpose and a limited number of questions helped ensure a clear line of sight towards action (reviews 122, 155).

The production of realistic and actionable recommendations was one of the most important factors in determining use of reviews. Of 24 online survey respondents, 22 identified the quality of the recommendations as being important to whether a review was used or not; and 18 of these respondents identified this as a *very* important factor. This finding was echoed in four of the five case studies. Case study respondents indicated that to have a high use, reviews needed to be practical and include realistic actions that resonate with commissioning body and donors (reviews 16, 184, 155).

A key challenge that review teams have faced in generating practical and actionable recommendations was when they had been asked to ***conduct reviews of projects that***

were no longer aligned with the future plans and strategy of the aid administration. This occurred in the East Timor and Pakistan case studies (reviews 244, 184). In both examples, changing aid priorities rendered both reviews largely irrelevant. In East Timor for example, the embassy already had decided to close down the project before the end review started. The email survey also revealed a number of cases in Vietnam, Sri Lanka and Malawi where changing donor priorities limited the use of reviews.[60]

Planning and delivering reviews in a consultative way is a key factor in determining use. In four of the five case studies the Norwegian aid administration emphasised close collaboration with the grant recipient at all stages of the review process – TOR, planning, implementation and follow-up – was important to providing the foundations for use or helped explain why there had been limited use (reviews 184, 155, 244, 122). In the East Timor case study for example, there was low buy-in to the review process from the grant recipient's side and the review was seen as an internal matter for the embassy that was not relevant for them. Conversely, in the Madagascar case study, a high level of stakeholder engagement, and the fact that the grant manager led the review, underpinned wide use of the end review findings.

The timing of a review impacted upon use. Ensuring a review was completed at the right time to feed into a decision was central to its utility. Some 21 of 24 respondents to the online survey felt this was important or very important to use. The case studies confirmed that timing was key to use and it was clear that the reviews for which use was highest had been commissioned to coincide with key decisions and/or moments in the implementation cycle. In these three cases, the timeliness of the reviews contributed to high use. In the Madagascar case, the review was timed so that it fed directly into the discussions of the next funding period (review 122). Similarly, in the Nepal case, the review was conducted at the same time as the government started to revise their energy policy (review 155). The Benguela end review coincided with the end of a cooperation phase with a partner and helped inform the follow-up (review 16).

60 In the case of the 'Enhancing Capacity to Control and Manage Biosafety and Biosecurity in Vietnam', the response notes that the 'grant management portfolio at the embassy in Hanoi has been reduced to only a few remaining development projects with final disbursement in 2016. So the utilisation of the lessons learned at this Embassy would be limited'. In the case of the 'Project for Rehabilitation through Education and Training Opportunities for Training in Needed Skills in Sri Lanka', the respondent notes that though the review gave some valuable input for a new phase, the 'Embassy however is not part of the continuation of funding due to change in priorities.' Similarly, for the 'Mid-Term Review for Lake Chilwa Basin Climate Change Adaptation Programme', the respondents note that 'we should keep in mind that the Embassy is currently reducing the number of agreements and as a result the planned phase II of support to Lake Chilwa will not be supported'.

6. Conclusions

The evaluation has drawn together findings from four components relying on quantitative and qualitative data analysis to answer the evaluation questions. Notwithstanding the limitations noted in Chapter 4, the analysis has allowed an integrated assessment of what factors determined quality and use in the 2014 sample examined. The approach has delivered several overarching, reliable conclusions.

1. The quality of reviews varied across the different quality criteria. Reviews tended to score well on how they defined their rationale, purpose and scope, and on answering the evaluation questions posed in the TOR, and in producing a useful set of recommendations. On other hand, reviews scored poorly in terms of stating their limitations, in addressing the ethics of how they conducted their work, and in setting out their analysis so that the findings were properly justified. Above all, two thirds of reviews had an inadequate rating for methodology, and did not set out how the data used to produce findings and recommendations were collected or analysed.

2. TORs were rated well in how they set out the rationale and purpose, the scope and questions to be evaluated, as well as the review process – all areas likely to be central in determining the direction of the subsequent review. They were rated less well in how they set out the programme logic, selected the review criteria, discussed the context and dealt with Norwegian aid's cross-cutting themes.

3. The quality of TORs was linked to the subsequent quality of the review report; a finding that emerged from the quality review, the case study and the online survey. In addition to the quality of the TOR, the level of resources allocated to a review was an important determinant of quality, as was the calibre of the review team. There was limited evidence to suggest that stakeholders within the Norwegian aid administration believed that a robust methodology was an important factor in review quality.[61]

4. The level of use of a review was influenced by the quality of the TOR and whether the review delivered against them, how realistic and actionable the recommendations were, whether key stakeholders had been engaged in the review process, and finally how well timed the review would fit with pending decisions related to the programme or project.

5. There was a serious gap in the provision of technical guidance for designing and managing reviews. The GMM was the only official source of guidance for commissioning and managing reviews, and grant managers often used material from other agencies or older Norad documentation. It is worth highlighting that the GMM stresses the importance of assessing the results of the projects, especially effectiveness and efficiency. Still, of the five OECD-DAC criteria included in the quality review, these are the two that had the lowest quality, and in particular two-thirds of the quality assessment sample did not address efficiency in line with international norms.

61 As requested in this evaluation's TOR, a set of best practice examples has been selected in Annex 10 covering TORs and reviews from the 2014 sample, and in addition the best practice cases mentioned by online survey respondents.

6. The evaluation found that decisions about Norwegian aid projects (whether under implementation or newly designed) were being taken on review findings and recommendations that were not always sound. Grant managers regarded reviews of being high quality when they were usable, by which they meant that findings and recommendations were actionable and targeted, and met the purpose set in the TOR. Even in the case where a review was of low quality, it seems that it was deemed of sufficient quality if it revealed something important and useful that could be implemented to improve the project. Yet there was less attention paid to whether the review was reliable or robust and met basic evidentiary standards for evaluations.

7. Although the 2014 reviews generally failed to be instruments for reliably documenting the results of Norwegian aid, they were in practice important management tools for the grant managers and units responsible for the grants. Review reports were highly used, and considered very useful in the aid administration. Use was mostly for instrumental purposes – to improve a project, to prepare a new grant or to feed into policies – which are important ways for reviews to be used. Nevertheless, whether reviews are used for these purposes or for conceptual use, there should be a sound information base that is used in an analytical way to draw conclusions. Finally, the administrative systems for providing efficient access to review documents within the administration have limitations and these hinder the wider use of reviews and corporate learning.

Annex 1 Terms of reference

THE QUALITY OF REVIEWS AND DECENTRALISED EVALUATIONS IN NORWEGIAN DEVELOPMENT COOPERATION

1. Introduction

Reviews and decentralised evaluations constitute an important part of the evidence base of Norwegian development cooperation. The annual government budget proposal for 2016 (Prop 1 S 2015–16) announced the introduction of stricter requirements to undertake reviews and evaluations in the project cycle. Recent studies suggest, however, that the quality of reviews and decentralised evaluations in Norwegian development cooperation is variable.[62] With increasing demand for evaluation to document the results of Norwegian development cooperation, there is need for more information about the quality of these evaluations, as well as a better understanding of factors contributing to quality and use of evaluation findings, conclusions and recommendations.

In preparation for this evaluation, the Evaluation Department commissioned a Mapping Study[63] in order to get a better overview of the extent of reviews and evaluations in Norwegian development cooperation. This is part of the data material for this evaluation, and is attached to the tender document as Annex 7.

2. Reviews and evaluations in Norwegian development cooperation

The obligation to evaluate government-funded efforts is established in the Regulations for Financial Management in the Government Administration.[64]

In the Norwegian aid administration,[65] evaluation is undertaken at several levels.

Centralised evaluations are undertaken by the Evaluation Department in Norad, which has a separate mandate[66] to initiate and carry out evaluations of Norwegian development cooperation. These centralised evaluations are normally more overarching thematically and geographically, and normally cover more than one programme/initiative. Around ten such evaluations are produced each year. These are not part of the scope of this evaluation.

Most evaluations of Norwegian aid are commissioned by the unit responsible for grant management (embassies, MFA, Norad),[67] implementing partners/grant recipients,[68] and other agencies/co-sponsors. In the Norwegian

62 Norad (2014) 'Can We Demonstrate the Difference that Norwegian Aid Makes?' Evaluation Report 1/2014, Norad, Oslo; OECD (2014), OECD Development Co-operation Peer Reviews: Norway 2013, OECD Publishing.

63 Norad Report 'Study of Reviews and Evaluations in Norwegian Development Cooperation – Mapping', draft October 2015.

64 'Reglement for økonomistyring i staten'(2003) and 'Bestemmelser om økonomistyring i staten' https://www.regjeringen.no/globalassets/upload/fin/vedlegg/okstyring/reglement_for_okonomistyring_i_staten.pdf Among several guides to supplement these regulations, the Minsitry of Finance has issued a guide for undertaking evaluations 'Veileder til gjennomføring av evalueringer' (2005).

65 For this purpose, this includes The Ministry of Foreign Affairs, Royal Norwegian Embassies managing ODA funds and Norad. Norfund and FK Norway, formally part of the Norwegian aid administration, are not part of this review.

66 https://www.norad.no/globalassets/filer-2015/evaluering/evaluation-instructions-from-23.-november-2015.pdf

67 Norad, in line with its mandate as quality assurer of Norwegian assistance, will also commission reviews on behalf of embassies and the MFA, as part of its technical support.

68 A large part of the Norwegian budget for ODA goes through the UN system and other multilateral organisations. Evaluation of these funds is through the evaluation systems of each of these organisations. Norwegian follow-up is mainly through participation in governing boards. This is not part of the scope of this evaluation. Programme support to UN-organisations at country level will be subject to reviews and evaluations where donors may be more involved in commissioning and carrying out the review. These reviews/evaluations are part of the scope of this evaluation.

aid administration, most of these initiative-level, *decentralised* evaluations are normally referred to as reviews, while the term evaluation is used for larger studies that are often broader in scope. The exact number of reviews and decentralised evaluations undertaken per year is not known, but the Mapping Study identified 235[69] reviews in the period January 2012 – May 2015, 60–70 per year.

Guidance for why, when and how to undertake reviews is given in the Grant Management Manual[70] (GMM) and requirements are specified in the rules[71] for each grant scheme. The GMM defines a review as '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). The GMM stipulates that the following factors should be addressed in a review: *'results in relation to the goal hierarchy (results framework) and implementation plans and budgets, as well as efficiency and effectiveness, risks, the capacity of the grant recipient and the models and methods employed in the project or programme'*. Further, it states that the following factors may be included: *'the effect of the project or programme in relation to external factors, the benefit achieved through changes in the operating conditions and the need and potential for reducing risk'*. In other words, the focus of these reviews is on operational aspects and factors influencing implementation. In the terminology of the OECD-DAC criteria for evaluating development assistance, 'effectiveness' and 'efficiency' are a primary concern, while criteria 'relevance' and 'sustainability' are less pronounced, though capacity and risk assessments may be elements of these criteria. Impact is also mentioned, though it is normally beyond the scope of such reviews.

The GMM states that cross-cutting issues in Norwegian development cooperation, should be taken into consideration in all interventions. In the period under evaluation, these were women's rights and gender equality; climate and environment; and anti-corruption. The GMM is not explicitly stating that reviews and evaluations should take these issues into account. But cross-cutting issues are seen as part of risk management, and as such, are among the aspects to be covered by a review. The Mapping Study found that reviews covered cross-cutting issues to a large degree, particularly gender equality (84% of reports).

The Mapping Study also indicated that 95% of the reviews addressed outcome. This evaluation will build on and expand the mapping, and assess the quality of findings and conclusions regarding effectiveness and other aspects.

Apart from the GMM and the grant scheme rules, there are currently no handbooks, standards, templates or guidance notes, nor any help desk or external quality assurance function to aid grant managers in commission-

69 The total number identified was 274, which included organisational reviews and thematic reviews, which are not part of the scope of this evaluation.

70 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).

71 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.

ing and managing reviews.[72] The Mapping Study found indications of low awareness regarding formal requirements of such reports, in that TORs were attached to only 60% of the reports. Furthermore, reviews could not easily be traced to the right project/programme agreement, as reports and TORs lacked reference to the agreement number or other programme/project identification in many cases.

In the period under evaluation (2014), grant scheme rules had few fixed requirements in terms of which interventions should be subject to a review. The grant scheme rules specified that the decision to conduct a review should be at the discretion of each grant manager, based on an assessment of the risk and significance of the intervention in question. Grant scheme rules for bilateral development cooperation[73] were a notable exception, wherein reviews were mandatory for agreements above a threshold of NOK 50 million. The Ministry of Foreign Affairs issued new grant scheme rules in February 2016, making reviews mandatory for programme/project agreements of a duration of over two years, and for agreements above a certain financial threshold, depending on the grant scheme. It is not clear how this will affect the number of reviews commissioned each year within the various grant schemes.

Grant managers are responsible for commissioning and carrying out reviews, as well as for follow-up of review findings and recommendations. Determining utility and use of reviews in the management cycle is therefore relevant.

Beyond this operational use, many reviews and decentralised evaluations will probably contain analyses, lessons, recommendations and information about the results of Norwegian development cooperation that is useful outside the programme under review. Some may also present general lessons learned. In the current set-up, reviews and decentralised evaluations do not effectively feed into an evidence base of Norwegian development cooperation.

Responsibility for tracking and collecting reviews done throughout the aid administration is not clear. The grant management system (PTA) of the Norwegian aid administration has a report function in place to track reviews, and the GMM has a requirement to register planned and completed reviews in PTA. This represents a potential source of credible information about the extent of reviews and evaluations. However, it is not kept updated by all grant-managing units. Therefore, it is not known how much of the annual development cooperation budget is subject to a review, and whether it is the most significant programmes that are reviewed.[74]

Reviews may be published at norad.no as part of the report series Norad Collected Reviews, though this is not a requirement in the GMM or the grant scheme rules. New grant scheme rules, as of February 2016, stipulates that

72 Resources that are frequently used, but do not have any official status in the current system, are SIDA Evaluation Manual 'Looking Back Moving Forward' (2004) and the former grant management manual, the Development Cooperation Manual (Norad, 2005), as well as various OECD-DAC evaluation resources and some former guidelines and handbooks produced by Norad on specific issues.

73 Regional allocation, Budget Ch. 150.78.

74 The Mapping Study, given its restricted scope, was not able to consistently identify the chapter/post of the programmes/projects under review.

reports be submitted to the Evaluation Portal, managed by the Norwegian Government Agency for Financial Management, in which all evaluations by government agencies should be registered. This should ensure an overview of reviews and evaluations in the future.

3. Quality of reviews and evaluations

In line with the purpose and scope of this evaluation, a relatively broad understanding of quality will guide the analysis and assessment, including aspects of utility, timeliness and relevance of reviews, in addition to technical quality (cf. the OECD-DAC quality standards).[75] Assessment of reviews should take into account their restricted scope as well as their particular purpose, which is normally programme-specific, and by nature inextricably linked to use of findings, conclusions and recommendations. It should be noted that the requirements and expectations for the quality of reviews in Norwegian aid management are lighter than what is normally the case for evaluations, e.g. expressed in the OECD-DAC quality standards, although the same principles should apply.

The evaluation team will develop the quality assessment criteria during the inception phase, in consultation with the Evaluation Department.

A few studies have looked into aspects of quality and use of reviews and evaluations of Norwegian aid.[76] They found weaknesses at different levels: Inadequate analysis of results achievement and causal relationships; lack of discussions of limitations of data material, and of the strengths and weaknesses of the analytical approach; weak logical link between findings and conclusions. Resources made available for reviews were found to be quite limited. Still, there is no consensus across studies: overall conclusions on the quality of reviews vary from 'by and large satisfactory' to 'generally poor'. In terms of utility and use, the finding from the study of NGO-commissioned reviews, was that 'instrumental' and 'process' use of reviews is strong, but that 'conceptual' use is weak, implying less attention to general lessons learned.[77]

Other donors have reviewed their decentralised evaluations and found varying quality.[78] Findings on factors contributing to quality are similar across many of these studies: evaluation team skills, resources allocated to evaluation, clarity of purpose, the number of evaluation questions, capacity (time and skills) of the commissioner, institutional factors and the quality of monitoring data.[79] The importance of these factors is not new and they are likely to apply also in the case of reviews/decentralised evaluations in the Norwegian aid administration.

75 http://www.oecd.org/dac/evaluation/qualitystandardsfordevelopment-evaluation.htm

76 Norad Report 1/2014; OECD-DAC Peer Review 2013; Norad Report 7/2012 'A Study of Monitoring and Evaluation in Six Norwegian Civil Society Organisations'; Norad Report 8/2012 'Use of Evaluations in the Norwegian Development Cooperation System'.

77 Norad Report 7/2012.

78 Examples include: DFAT (2014) Quality of Australian aid operational evaluations. Office of Development Effectiveness, Department of Foreign Affairs and Trade, Australian Government; SIDA (2008) Are SIDA Evaluations Good Enough? An assessment of 34 Evaluation Reports. Forss et. al. Sida Studies in Evaluation 2008:1.

79 CDI Practice Paper 09 March 2015 'Improving Quality: Current Evidence on What Affects the Quality of Commissioned Evaluations'. Rob Lloyd and Florian Schatz.

4. Purpose and objectives

The overall purpose of this evaluation is to contribute to good quality reviews and decentralised evaluations in Norwegian development cooperation. Main intended users are the Ministry of Foreign Affairs, Norwegian embassies managing ODA funds, Norad and other parts of the aid administration.

The evaluation will serve as input into a discussion on the organisation and management of decentralised evaluation and reviews in the Norwegian development administration.

The objectives of this evaluation are to:

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

2. Examine the use of review findings, conclusions and recommendations; and

3. Identify factors contributing to quality and use of reviews and decentralised evaluations in Norwegian development cooperation.

5. Scope and evaluation object

The evaluation object is reviews and decentralised evaluations commissioned by the Norwegian aid administration (Norwegian embassies, the Ministry of Foreign Affairs and Norad), or in which it has taken active part.[80]

The evaluation will cover 74 reviews/decentralised evaluations finalised in 2014 and identified through the Mapping Study.[81]

For this purpose, a review or evaluation should be understood not just as the end-product/final report, but also include the process of undertaking a review from the perspective of the commissioning body, including the terms of reference.

'Reviews' in this context include mid-term reviews, end reviews and decentralised evaluations of programmes or interventions. Organisational reviews and thematic reviews fall outside the scope of this evaluation.

6. Evaluation questions

The following questions will guide the evaluation:

1. What are the main strengths and weaknesses of reviews and decentralised evaluations of Norwegian development cooperation?

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

3. From the perspective of stakeholders, to what extent are reviews timely, and present relevant and realistic recommendations?

80 Though logically part of the universe of this evaluation, reviews commissioned by NGO grant recipients are excluded from the scope. The quality of these reviews has been assessed in two previous evaluations, Norad Report 1/2014 and Norad Report 7/2012. Evaluations of UN and other multilateral organisations are not part of scope, except decentralised evaluations at country level commissioned or co-sponsored by Norwegian embassies.

81 The year 2014 has been selected because it can expected to be the most complete of the years covered in the Mapping Study (2012–15).

4. To what extent have review findings, conclusions and recommendations been used by the unit responsible for managing the grant to the intervention that has undergone review?

5. What are the main factors contributing to quality and use of reviews and decentralised evaluations?

6. To what extent do reports present any general lessons learned with relevance beyond the intervention under review?

During the inception phase, the evaluation team will develop and elaborate the questions in line with the objectives and approach outlined in these TORs, and in consultation with the Evaluation Department.

7. Methodology

All parts of the evaluation shall adhere to recognised evaluation principles and, where relevant, the OECD Development Assistance Committee's quality standards for development evaluation, as well as relevant guidelines from the Evaluation Department (available at norad.no/evaluationguidelines).

In the inception phase, the evaluation team will develop the analytical framework and criteria for the quality assessment, in consultation with the Evaluation Department. The framework will build on the requirements for reviews as expressed in guidance documents for grant managers and technical advisers commissioning (particularly the GMM) on the one hand, and accepted international standards, such as the OECD-DAC quality standards for development evaluation, on the other. Relevant quality aspects include clarity of purpose of the review, reports' methodological and analytical soundness, utility, relevance and timeliness. The analysis may also include an assessment of the credibility of report findings and conclusions on efficiency, effectiveness, relevance, sustainability, and (if applicable) impact.

The evaluation team will propose the methodological approach, which may include the following components:

Desk review: The evaluation team will undertake a desk review of the reviews and decentralised evaluations finalised in 2014, identified in the Mapping Study. In addition to the final reports, the document review shall include the TORs document and – for a sub-sample – inception reports, work plans or other relevant documentation, in order to explore factors contributing to quality and use of reviews and decentralised evaluations.

Case studies: A smaller sample of reviews and decentralised evaluations should be analysed to determine the use of findings, conclusions and recommendations in grant management. In addition to interviews with grant managers responsible for follow-up of the reviews/evaluations, and possibly a survey, this will include document review of relevant appropriation documents, agreements, memos from annual meetings, or other documentation of dialogue with partners regarding follow-up of reviews.

Interviews: It will also include interviews and possibly a survey among stakeholders responsible for commissioning and follow-up of

selected reviews, the MFA, Embassies, Norad and partners where relevant, as well as consultants having carried out the reviews/ evaluations (covering both internal and external teams).

The data collected shall be supplemented/ triangulated with data from other relevant primary and secondary sources.

As part of the evaluation, the evaluation team will identify of a limited number of best practice examples of evaluation products. This may include TORs documents, inception reports, final reports or sections of final reports, such as the recommendations section or a general lessons section, and present them as an annex to the evaluation report, or other appropriate presentation format, including a brief description of their particular strengths.

The team can propose to remove any review from the sample that, during the initial analysis of the material, proves not to fit the scope of the evaluation. The material may contain a few reports (e.g. self-evaluations, final narrative reports etc.) that fall outside the scope. The evaluation team may propose an alternative approach that responds to the purpose and objectives in this TORs in other ways than those laid out above, demonstrating comparable rigour and ability to respond to the evaluation questions.

8. Organisation

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

All decisions concerning the interpretation of these TORs, and all deliverables are subject to approval by the Evaluation Department.

The team is entitled to consult widely with stakeholders pertinent to the assignment. Access to archives and statistics will be facilitated by Norad and stakeholders.

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

9. Deliverables

The deliverables in the consultancy consist of the following outputs:

- Draft inception report, including framework for quality assessment of reviews/decentralised evaluations – to be approved by the Evaluation Department
- Final inception report
- Draft final report not exceeding 40 pages, excluding summary and annexes, for preliminary approval by the Evaluation Department and circulation to the stakeholders. After circulation to the stakeholders, the Evaluation Department will provide feedback
- Best practice examples of evaluation products, to be submitted with the draft report
- Final evaluation report
- Seminar/workshop in Oslo to present the final report
- Evaluation brief not exceeding three pages.

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). The Evaluation Department retains the sole rights with respect to all distribution, dissemination and publication of the deliverables.

Annex 2 References

Australian Department of Foreign Affairs and Trade, *Quality of Australian Aid Operational Evaluations*, Office of Development Effectiveness, Canberra, June 2014

Barr, J. (Itad), Rinnert, D. (DFID), Lloyd, R. (Itad), Dunne, D. (DFID), Henttinen, A. (DFID), *The Value of Evaluation: Tools for Budgeting and Valuing Evaluations*, Discussion Paper, DFID, 2016

Chelimsky, E. Integrating Evaluation Units into the Political Environment of Government: The Role of Evaluation Policy, in Trochim, W.M.K., Mark, M.M. and Cooksy, L.J. (eds.) *Evaluation Policy and Evaluation Practice. New Directions for Evaluation*, 123: 51–66, 2009

Cooksy, L.J. and Mark, M.M. Influences on Evaluation Quality, *American Journal of Evaluation* 33.1: 79–84, 2012

Global Affairs Canada, Meta-evaluation of Global Affairs Canada's Decentralised Evaluations: FY 2009/10–13/14, Ottawa, due for publication late 2016

Johnson, K., Greenseid, L.O., Toal, S.A., King, J.A., Lawrenz, F. and Volkov, B. Research on evaluation use a review of the empirical literature from 1986 to 2005. *American Journal of Evaluation*, 30(3), 377–410, 2009

Laast, de B., Evaluator, Evaluand, Evaluation Commissioner, a Tricky Triangle, Ch. 2 in Loud and Mayne (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*, Thousand Oaks, CA: Sage, 2014

Lloyd, R. and Schatz, F. *Improving quality: current evidence on what affects the quality of commissioned evaluations*, Centre for Development Impact Practice Paper, No 9, March 2015, IDS; 2015

Loud, M. and Mayne, J. (eds) *Enhancing Evaluation Use: Insights from Internal Evaluation Units*, Thousand Oaks, CA: Sage, 2014

Management Systems International, Meta-evaluation of Quality and Coverage of USAID Evaluations 2009–2012

Ministry of Foreign Affairs Norway, *Grant Management Manual*, Management of Grants by the Ministry of Foreign Affairs and Norad, 2013

Norad, *Can We Demonstrate the Difference that Norwegian Aid Makes?* Evaluation of Results Measurement and How This Can be Improved, Oslo, 2014

NORDIC Consulting Group, Study of Reviews and Evaluations in Norwegian Development Cooperation – Mapping, Final Report, October 2015

Organisation for Economic Cooperation and Development, *Quality Standards for development evaluation*, DAC Guidelines and Reference Series, Development Assistance Committee, 2010

Schwandt, T. Defining 'Quality' in Evaluation. *Evaluation and Program Planning*, 13(2): 1990

Stetler, C.B. Ch. 3: Stetler Model. In J. Rycroft-Malone and T. Bucknall (eds), *Models and frameworks for Implementing Evidence-Based Practice: Linking evidence to action*. Evidence-based Practice Series. Wiley-Blackwell, Oxford, 2010

UN Women, Global Evaluation Reports Assessment and Analysis System (GERAAS), 2014

UNDP, *Annual Report on Evaluation 2013*, New York: Independent Evaluation Office, 2013

UNICEF, The Global Evaluation Report Oversight System (GEROS) 2010–15, 2015

USAID, *Meta-evaluation of Quality and Coverage of USAID Evaluations 2009–12*, prepared by Management Systems International. UNDP, Annual Report on Evaluation, Independent Evaluation Office, UNDP, 2013

Winckler Andersen, O. *Some thoughts on development evaluation processes*, IDS Bulletin 45(6): 77–84, 2014

Yin, R,K. *Case Study Research: Design and methods,* 5th edition, Thousand Oaks, CA: Sage, 2014

Annex 3 Methodology

1 Email survey of grant managers

The first component was an email survey of grant managers/project officers responsible for the reviews in 2014 to gather views on the quality and use of the report, and to obtain additional documentation and details of the review. The survey was a light-touch exercise with a limited set of questions, in an effort to allow busy aid officials to respond in a timely fashion. The email was issued in advance of the evaluation inception report because of the urgent need to contact staff before the July leave period. A total of three reminders were sent over a period of one month in an effort to bolster the response rate. The evaluation team also followed up directly with respondents who had specific requests or who had inadvertently sent the wrong documentation.

The sample for the email survey was drawn from the 74 reviews published in 2014 by MFA, Norwegian embassies, Norad and partners that fall within the scope of this evaluation. Four studies were excluded from the set of 74 for various reasons[82] leaving a remaining sample of 70 reviews. The evaluation team sent an email survey to the 60 staff responsible for commissioning the 70 reviews (grant managers/project officers) across the relevant agencies (MFA, Norad, embassies). A total of 35 replies were received and analysed (see Annex 4).

2 Quality assessment

The second component was a quality assessment of reviews of Norwegian development cooperation and associated TORs. The quality assessment was undertaken using a template that focused on factors that were most associated with overall review quality and use (to the extent that this could be judged from the TOR and review documentations alone). The quality areas chosen for these assessment templates drew on two main references: (1) the OECD-DAC standards for evaluation; and also (2) Norad priorities in terms of cross-cutting issues. The quality areas and the related quality standards that were used to guide the review team's assessment are provided in Annex 6, Appendix 1.

The template consisted of two quality assessments tools: (1) for the TOR, and (2) for the review reports. The reviewer assigned ratings based on the documentary evidence available, providing a justified rating for a series of quality areas (e.g. evaluation purpose, evaluation scope, data analysis, etc.) against the descriptions of satisfactory quality for those areas. In each case, the rating reflected the degree of confidence that the evaluation documents provided the reviewer with regard to the issue covered by a quality area as indicated in 3.1 of this report. There was also an option for the reviewer to note if a particular quality area was not applicable because it was not relevant to the review.

Prior to conducting the quality review a rigorous piloting of the template took place to ensure rating consistency across the team. In total,

82 One was found to have been published in 2013, one was in Portuguese (and the team does not have this language among its members), one was a study conducted by Itad and so has been removed to prevent a conflict of interest, and one was a case where the initiative was funded by an umbrella organisation for Norwegian mission organisations, Digni, and so it falls under the category grant recipients' reviews, which is not part of the scope.

the evaluation team conducted a pilot of three reviews for this purpose. For each pilot, a conference call was organised to allow the evaluation team to discuss discrepancies in ratings. Through this process, consensus was developed on quality areas and their rating across the different sections of the template. It also allowed for the template to be further refined with suggestions for improvement being signed off by Norad.

The sample was drawn from the 74 reviews published in 2014 by MFA, embassies, Norad and partners that fall within the scope of this evaluation. However, based on the evaluation resources and the limited timeframe to complete the quality review, agreement was reached with Norad to limit the sample to 60 cases. An initial sample was therefore defined using a systematic random sampling procedure to select 60 cases.[83] A remaining 10 cases were held in reserve to be used as replacements in cases where, following the initial staff survey and request for all relevant documents, key documentation (especially the TOR) was still missing from the initial sample. Using the data being collected through the email survey, the evaluation team identified a number of reviews where additional documents could not be retrieved because of embassy closures and staff rotation. Also, two of the reviews from the sample of 60 had not been commissioned by the Norwegian aid administration and had to be replaced. In total, eight reviews were replaced using purposive sampling to ensure a geographically balanced sample to the greatest extent possible. The final sample of 60 reviews can be found in Annex 5.

3 Case studies

The aim of the third component was to provide an in-depth assessment of a sample of reviews together with their associated documents such as TORs, inception reports and management responses, analysing enablers and barriers of review quality and use.

The approach was based on understanding evaluation as a process. Evaluation reports are only one product of this process, which includes the stages of planning, implementation, reporting and use. Evaluation quality cuts across all stages of the evaluation process and needs to be assessed within each phase.[84] The approach recognised that the evaluation process is embedded in the relationship between the evaluation commissioner and the evaluation team and their respective capacities, and the wider institutional environment in which the evaluation is being conducted.[85] While the quality reviews served to better understand the reporting phase, the case studies focused on the three other phases of an evaluation: planning, implementation and use. The approach assessed quality in these three different phases through a set of common questions. Annex 7 details the methodology of the case study approach.

83 A sampling interval is obtained by dividing the universe by the required sample (70/60 = 1.2 approx.). Then taking a random start point between 1 and 1.2, every xth case + the required interval (converted to an integer) is selected. This will ensure that every evaluation has a known and equal chance of being selected, and will ensure that the 60 cases are distributed across the pool of evaluations.

84 Lloyd and Schatz (2015) op. cit.

85 Winckler Andersen, O. (2014) Some thoughts on development evaluation processes, *IDS Bulletin* 45(6): 77–84.

Five case studies were purposefully selected to represent different levels of quality and use. To obtain cases of high and low quality and use, the evaluation drew on evidence obtained from the quality review and the email survey. From a set of ten reviews proposed, the final five choices were agreed with the Norad Evaluation Department. These choices also reflected a balance of regions and of sectors that were of wider importance in Norwegian development cooperation.

Data collection was undertaken by two evaluation team members through phone interviews. Three types of respondent were contacted: the grant manager, a review user in the Norwegian aid administration other than the grant manager, and the consultant team leader who led the review. Each interview lasted up to one hour and followed a semi-structured questionnaire. Interview guides for each stakeholder group can be found in the Annex 7. The questions developed reflect a deductive approach that draws on existing literature on evaluation quality[86] and use,[87] including findings on key factors influencing evaluation quality in recent meta-evaluations such as DFAT (2014), Norad; Itad/Chr. Michelsen Institute (2014); USAID (2013), UNDP (2013).[88]

4 Online survey

To gather wider perceptions within the Norwegian aid administration, and to test emerging findings from components 1, 2, and 3 among a wider sample of staff, an online survey was conducted. The online survey was also an exercise in eliciting staff views on the quality and use of reviews, and on the support available to commission and implement them.

The online survey consisted primarily of closed-ended questions with some opportunities for the respondent to provide open-ended responses. The tool was piloted with the Evaluation Department, a member of staff from a Norwegian embassy, and internally within the evaluation team to ensure consistency in language and the flow of the survey. The survey was refined on a number of occasions before a final version was signed off by Norad. The implementation of the survey was through the online survey tool Survey Monkey.[89] This helped to ensure that the invitation email to complete the survey guaranteed anonymity and that responses were handled securely. An advisory was sent by the Norad Evaluation Department prior to distributing the survey. This was followed by a reminder before a third and final reminder was sent with an accompanying request from the Evaluation Department to complete the survey. For each reminder,

86 Cooksy, L.J. and Mark, M.M. (2012) Influences on Evaluation Quality, *American Journal of Evaluation* 33(1): 79–84. Chelimsky, E. (2009) Integrating Evaluation Units into the Political Environment of Government: The Role of Evaluation Policy, in Trochim, W.M.K., Mark, M.M. and Cooksy, L.J. (eds) *Evaluation Policy and Evaluation Practice. New Directions for Evaluation,* 123: 51–66.

87 Johnson, K., Greenseid, L.O., Toal, S.A., King, J.A., Lawrenz, F. and Volkov, B. (2009) Research on evaluation use a review of the empirical literature from 1986 to 2005. *American Journal of Evaluation*, 30(3): 377–410

88 Australian Department of Foreign Affairs and Trade, Quality of Australian Aid Operational Evaluations, Office of Development Effectiveness, June 2014; Itad/Chr. Michelsen Institute (2014) *Can We Demonstrate the Difference that Norwegian Aid Makes?* Evaluation of Results Measurement and How This Can be Improved, Oslo: Norad; USAID (2013) *Meta-evaluation of Quality and Coverage of USAID Evaluations 2009–2012*, prepared by Management Systems International. UNDP (2013) *Annual Report on Evaluation 2013*, New York: Independent Evaluation Office, UNDP.

89 www.surveymonkey.net

those who had already responded to the survey were removed from the reminder list.

The intention of the online survey was to broaden the evaluation knowledge base, and it was therefore relevant to include in the sample staff beyond those already contacted in the email survey. It was also important to reflect a cross-section of grant commissioners and users in MFA, Norad and embassies, and to include relevant managers, advisers and programme staff who were in positions and sections that were involved in reviews. Rather than send out a blanket email to all staff in these agencies, the aim was to select the most relevant people in them: those that were likely to have experience of either commissioning or using the reviews that the evaluation focuses on. In the case of MFA, staff from the Regional Department and from the UN and Humanitarian Department were therefore targeted. In the latter, the selection was from only the humanitarian and democracy sections. In the case of Norad, staff from the four thematic sections and from communications were selected. For the embassies, the choice was from focus countries and others with a larger staff complement working in development cooperation.[90]

Based on the above, a final universe of some 300 staff to be sampled was obtained from online portals and other sources in consultation with the Evaluation Department. From these names, a random sample of 120 staff was drawn, consisting of 40 staff from Norad, 40 from the MFA and 40 people from embassies. The sample size was chosen to be keep the survey focused on a smaller group that would be more likely to respond yet provide a sample large enough to give significant results. The survey was open for two weeks from 14–30 October. By the cut-off date, a total of 34 responses were received giving a response rate of 28%.

90 A paper produced by the study team is available that sets out the sample approach.

Appendix 1 Review templates

KEY INFORMATION

Key information					
				Eval Ref #	
Itad reviewer					
Project #	251				
Project Report title	Oil for Development Uganda 2009-2014: Review of Norway's Support to the Petroleum Sector in Uganda				
Project budget (Overall) NOK'000					
Region	Africa South of Sahara				
Country	Uganda				
Commissioner	Embassy				
Target Area	322				
Implementing Partner	Oslo Centre				
Project officer					
Unit	Emb. in Kampala/Uganda				
Type of evaluation	End review				
Evaluation Team	External				
Evaluation budget	Currency		NOK	Amount	
Evaluation days	Total Days		Fieldwork days		
Evaluation team	Total			Female	
	International			National	
Reviewed documents					
		Ratings	**Text**		
ToRs	*Completion*	0%	0%	Year	2014
Evaluation report	*Completion*	0%	0%	Year	2014

Key quality areas	Quality statement	Rating 1–4	Evidence and justification of the rating
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: · 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 ToRs contain a brief description of the context of the intervention being evaluated. This may include: · 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	The ToR states whether previous reviews exist, and If applicable, identifies relevant issues		
1.5 Object of the Review	The development intervention being reviewed (the review object) is clearly described, including: · period		
	· budget		
	· geographical area		
	· Intervention logic/theory of change/logic model		

Key quality areas	Quality statement	Rating 1–4	Evidence and justification of the rating
	· expected outcomes		
	· stakeholders		
	· organizational set-up		
1.6 Scope	The ToRs clearly define what will and will not be covered by the review, including: · What aspect/dimensions of the intervention		
	· the time period		
	· the geographic coverage		
1.7 Review criteria	Based on the review mandate, the ToR identifies the relevant criteria (OECD/DAC, cross-cutting themes and issues) for the review: · OECD/DAC: relevance, efficiency, effectiveness, impact and sustainability		
	· Cross-cutting themes: 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 TOR is feasible given the timeframe and resources provided		
	The ToRs 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 1–4	Evidence and justification of the rating
2. Review Process and QA			
2.1 Review process	The review ToR clearly explains what is expected of the Consultant in terms of: 1. having an inception stage		
	2. data collection and validation		
	3. preparing the review or review report		
	Roles and responsibilities of the team members (consultants) and of Norad/ MFA/Embassy/ Partner (review manager) are defined and appropriate to the review objectives		
2.2 Deliverables	The review ToR identifies the mandatory deliverables and milestones: · 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 ToRs specify that the review will follow professional norms and standards, including OECD/DAC.		
	Provisions for quality assurance mechanisms are included in the ToRs		
3. Overarching and cross-cutting criteria			
3.1 Gender	Gender dimensions and women's rights are explicitly addressed in all relevant parts of the ToRs (context, questions, approach, design, methods, team composition)		
3.2 Climate and Environment	Climate and environment dimensions are reflected in the TOR where appropriate (context, design, questions around effectiveness and impact)		
3.3 Anti-corruption	Anti-corruption issues are reflected in the TOR (e.g as part of risks or context)		

→

Key quality areas	Quality statement	Rating 1–4	Evidence and justification of the rating
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 ToRs	The ToRs provide a sound basis for the review, that will guide the review manager and team on how to fulfill effectively the objectives of the review		
List any examples of good practice			

REVIEW TEMPLATE

Key quality areas	Quality statement	Rating 1–4	Evidence and justification of the rating
1. Summary and style			
1.1 Executive summary	The review report contains an executive summary		
	The summary is complete and concise. It provides an overview of the report, highlighting the rationale, purpose and specific objectives of the review, the intervention, the scope of the review, the methodology used and the main findings, conclusions, recommendations and lessons of the review		
1.2 Style and structure	The structure of the report allows for a clear and logical flow of information from beginning to end.		
	The report is well written and properly edited		
2. Review purpose, objectives, object and scope			
2.1 Rationale and purpose of the review	The rationale, purpose, intended users and intended use of the review are stated clearly, addressing: · Why is the review 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	The specific objectives of the review clarify what the review aims to find out.		
	Any modification to the specific objectives stated in the ToR is explained		
2.3 Context of the development intervention	The review report describes the context of the development intervention, including: · policies, objectives and strategies of implementers		
	· development context, including socio-economic, political, cultural factors		
	· Key issues pertaining to Norway's cross-cutting themes (women's rights and gender equality; climate and environment; and anti-corruption) where applicable		

→

Key quality areas	Quality statement	Rating 1–4	Evidence and justification of the rating
2.4 Review object	The description of the intervention includes: · the time period, budget, geographical area		
	· components of the intervention		
	· expected outcomes;		
	· stakeholders		
	· organizational set-up/implementation arrangements		
	· discrepancies between the planned and the actual implementation of the development intervention are explained.		
2.5 Scope	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 ToR are explained.		
2.6 Review criteria and questions	The review should apply the agreed DAC criteria for evaluating development assistance (relevance, efficiency, effectiveness, impact and sustainability) and Norway's cross-cutting themes of gender equality, climate and environment, and anti-corruption) unless alternative criteria and questions are clearly defined in the ToR.		
	The review questions address all the review criteria and cross-cutting themes adequately.		
	Questions are clear, specific, and answerable.		
	Any modifications from the criteria and questions presented in the ToRs are explained and justified.		
2.7 Previous reviewss	Key findings and recommendations stemming from previous reviewss that have informed the current reviews are mentioned		

Key quality areas	Quality statement	Rating 1–4	Evidence and justification of the rating
3. Methodology			
3.1 Description of the design	The report describes: - the review approach (conceptual framework)		
	- the review design		
3.2 Sources of evidence	The review report describes: - the sources of information used (documents, respondents, administrative data, literature, etc.)		
	- sampling strategy		
3.3 Description of methods	The review (inception) report describes: - instruments/techniques used for data collection, including those used to collect gender-sensitive data and information.		
	- data analysis methods, including analysis of gender-sensitive data and information.		
3.4 M&E	The adequacy of M&E data/systems are described		
	The review makes use of the existing M&E data		
3.5 Methodological appropriateness and robustness	The selected review methodology (including approach, design, methods for data collection, analysis and sampling) is appropriate given the review purpose, objectives and approach and well justified.		
	Methods are linked to and appropriate for each review question.		
	Multiple lines of evidence are used.		
	The review cross-validates the information sources and assesses the validity and reliability of the data. Use of Triangulation in gathering evidence is sufficient		

Key quality areas	Quality statement	Rating 1–4	Evidence and justification of the rating
3.6 Limitations and challenges	The review report explains any limitations in process, data sources and sampling/samples, data collection and data analysis are explained as well as their implications in terms of validity and reliability.		
	Limitations regarding the representativeness of the sample for interpreting review results are explained.		
	Any obstruction of a free and open review process which may have influenced the findings is explained		
3.7 Ethics	Ethical issues such as privacy, anonymity, do-no-harm, inclusion/exclusion, and cultural appropriateness are described and addressed.		
	Ethical safeguards are described and 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)		
4. Application of selected OECD DAC criteria (where relevant from TOR)			
4.1 Relevance	The report correctly interprets and assesses relevance in the context if the initiative. It should refer to the extent to which the aid activity is suited to the priorities and policies of the target group, recipient and donor.		
4.2 Effectiveness	The report correctly interprets and assesses effectiveness: the initiative is assessed as meeting or likely to meet its objectives, and 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 alternatives approaches would have produced the same results for less resoures.		
4.4 Sustainability	The report correctly interprets and assesses sustainability: whether the benefits of an activity 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: whether the initiative is likely to or has begun to attain its longer term goals beyond the life of the intervention		

Key quality areas	Quality statement	Rating 1–4	Evidence and justification of the rating
5. Analysis, data, findings, conclusions and recommendations			
5.1 Review questions answered	The review report answers all the questions detailed in the ToR for the review.		
	The questions from the ToR, as well as any revisions, removals or additions to these questions, are documented in the report to enable readers to assess whether the review team has sufficiently addressed the questions, including those related to cross-cutting themes, and met the review objectives		
5.2 Programme logic	Is the ToC/programme logic assessed in a comprehensive manner, are any gaps identified and is it assessed against existing literature/evidence? Is a description of the assumptions underlying the ToC/programme logic 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 on results clearly distinguish outputs, outcomes and impacts (where approriate) and demonstrate the progression from implementation to results.		
	Attribution and/or contribution of the intervention to the result are discussed. There is an exploration of other factors which may have caused the results outside the intervention.		
5.5 Conclusions	Conclusions presents 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 well grounded in the evidence and follow logically from the conclusions.		

→

Key quality areas	Quality statement	Rating 1–4	Evidence and justification of the rating
5.7 Lessons learned	If present, lessons follow logically from the conclusions. Lessons should only be drawn if they represent contributions to general knowledge.		
	If not present, rate as N/R unless required by ToR in which case rate as 1.		
	Are there significant lessons of wider applicability or originality that contribute to the Norwegian aid administration's broader understanding beyond the intervention under review ? If there are, copy these into the text box D110.		
5.8 Integration of gender, climate and environment and anti-corruption	Gender dimensions (if requested in the TOR) inform the findings, conclusions, recommendations and lessons as appropriate. (if not then N/R)		
	Climate and environment issues (if requested in the TOR) are integated where appropriate into the findings, conclusions, recommenations and lessons.		
	Anti-corruption issues (if requested in the TOR) are integated where appropriate into the findings, conclusions, recommenations and lessons.		
OVERALL RATING of the Review			
Is this review an overall example of best practice? If not, are their particular elements that represent good practice within the review? If so list them			

Acronyms

BCC	Benguela Current Commission
CMI	Chr. Michelsen Institute
DAC	Development Assistance Committee
DFAT	Department for Foreign Affairs and Trade, Australia
GMM	Grant Management Manual
M&E	Monitoring and Evaluation
MFA	Ministry of Foreign Affairs, Norway
NOK	Norwegian Krone
Norad	Norwegian Agency for Development Cooperation
ODA	Official Development Assistance
OECD	Organisation for Economic Co-operation and Development
PTA	the acronym in Norwegian for the Norwegian aid Grant Management System
SIDA	Swedish International Development Agency
TOR	Terms of Reference
USAID	United States Agency for International Development
UNDP	Unit Nations Development Programme

Former reports from the Evaluation Department

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

2016

8.16 Country Evaluation Brief: Mozambique

7.16 Country Evaluation Brief: Afghanistan

6.16 Country Evaluation Brief: South Sudan

5.16 Evaluation of Norway's support for advocacy in the development policy arena

4.16 Striking the Balance: Evaluation of the Planning, Organisation and Management of Norwegian Assistance related to the Syria Regional Crisis

3.16 Real-Time Evaluation of Norway's International Climate and Forest Initiative. Literature review and programme theory

2.16 More than just talk? A Literature Review on Promoting Human Rights through Political Dialogue

1.16 Chasing civil society? Evaluation of Fredskorpset

2015

10.15 Evaluation of Norwegian Support to capacity development

9.15 Evaluation series of NORHED: Evaluability study

8.15 Work in Progress: How the Norwegian Ministry of Foreign Affairs and its Partners See and Do Engagement with Crisis-Affected Populations

7.15 Evaluation of Norwegian Multilateral Support to Basic Education

6.15 Evaluation Series of NORHED Higher Education and Research for Development. Evaluation of the Award Mechanism

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

4.15 Experiences with Results-Based Payments in Norwegian Development Aid

3.15 A Baseline Study of Norwegian Development Cooperation within the areas of Environment and Natural Resources Management in Myanmar

2.15 Evaluation of Norway's support to women's rights and gender equality in development cooperation

1.15 Evaluation of the Norwegian Investment Fund for Developing Countries (Norfund)

2014

8.14 Evaluation of Norway's Support to Haiti after the 2010 Earthquake

7.14 Baseline. Impact Evaluation of the Norway India Partnership Initiative Phase II for Maternal and Child Health

6.14 Building Blocks for Peace. An Evaluation of the Training for Peace in Africa Programme

5.14 Evaluation of Norwegian support through and to umbrella and network organisations in civil society

4.14 Evaluation Series of NORHED Higher Education and Research for Development. Theory of Change and Evaluation Methods

3.14 Real-Time Evaluation of Norway's International Climate and Forest Initiative: Synthesising Report 2007-2013

2.14 Unintended Effects in Evaluations of Norwegian Aid

1.14 Can We Demonstrate the Difference that Norwegian Aid Makes? Evaluation of results measurement and how this can be improved

2013

5.13 Real-Time Evaluation of Norway's International Climate and Forest Initiative: Measurement, Reporting and Verification

4.13 Evaluation of Five Humanitarian Programmes of the Norwegian Refugee Council and of the Standby Roster NORCAP

3.13 Evaluation of the Norway India Partnership Initative for Maternal and Child Health

2.13 Local Perception, Participation and Accountabillity in Malawi's Health Sector

1.13 A Framework for Analysing Participation in Development

2012

9.12 Evaluation of Norway's Bilateral Agricultural Support to Food Security

8.12 Use of Evaluations in the Norwegian Development Cooperation System

7.12 A Study of Monitoring and Evaluation in Six Norwegian Civil Society Organisations

6.12 Facing the Resource Curse: Norway's Oil for Development Program

5.12 Evaluation: Real-Time Evaluation of Norway's International Climate and Forest Initiative. Lessons Learned from Support to Civil Society Organisations

4.12 Evaluation of the Health Results Innovation Trust Fund

3.12 Evaluation of Norwegian Development Cooperation with Afghanistan 2001-2011

2.12 Hunting for Per Diem. The Uses and Abuses of Travel Compensation in Three Developing Countries

1.12 Mainstreaming disability in the new developmentparadigm

2012

9.12 Evaluation of Norway's Bilateral Agricultural Support to Food Security

8.12 Use of Evaluations in the Norwegian Development Cooperation System

7.12 A Study of Monitoring and Evaluation in Six Norwegian Civil Society Organisations

6.12 Facing the Resource Curse: Norway's Oil for Development Program

5.12 Evaluation: Real-Time Evaluation of Norway's International Climate and Forest Initiative. Lessons Learned from Support to Civil Society Organisations

4.12 Evaluation of the Health Results Innovation Trust Fund

3.12 Evaluation of Norwegian Development Cooperation with Afghanistan 2001-2011

2.12 Hunting for Per Diem. The Uses and Abuses of Travel Compensation in Three Developing Countries

1.12 Mainstreaming disability in the new development paradigm

2011

10.11 Evaluation of Norwegian Health Sector Support to Botswana

9.11 Activity-Based Financial Flows in UN System: A study of Select UN Organisations

8.11 Norway's Trade Related Assistance through Multilateral Organizations: A Synthesis Study

7.11 Evaluation: Evaluation of Norwegian Development Cooperation to Promote Human Rights

6.11 Joint Evaluation of Support to Anti-Corruption Efforts, 2002-2009

5.11 Pawns of Peace. Evaluation of Norwegian peace efforts in Sri Lanka, 1997-2009

4.11 Study: Contextual Choices in Fighting Corruption: Lessons Learned

3.11 Evaluation: Evaluation of the Strategy for Norway's Culture and Sports Cooperation with Countries in the South

2.11 Evaluation: Evaluation of Research on Norwegian Development Assistance

1.11 Evaluation: Results of Development Cooperation through Norwegian NGO's in East Africa

2010

18.10 Evaluation: Real-Time Evaluation of Norway's International Climate and Forest Initiative

17.10 Evaluation: Real-Time Evaluation of Norway's International Climate and Forest Initiative. Country Report: Tanzania

16.10 Evaluation: Real-Time Evaluation of Norway's International Climate and Forest Initiative. Country Report: Indonesia

15.10 Evaluation: Real-Time Evaluation of Norway's International Climate and Forest Initiative. Country Report: Guyana

14.10 Evaluation: Real-Time Evaluation of Norway's International Climate and Forest Initiative. Country Report: Democratic Republic of Congo

13.10 Evaluation: Real-Time Evaluation of Norway's International Climate and Forest Initiative. Country Report: Brasil

12.10 Evaluation: Real-Time Evaluation of Norway's International Climate and Forest Initiative (NICFI)

11.10 Evaluation: Evaluation of the International Organization for Migration and its Efforts to Combat Human Trafficking

10.10 Evaluation: Democracy Support through the United Nations

9.10 Study: Evaluability Study of Partnership Initiatives

8.10 Evaluation: Evaluation of Transparency International

7.10 Evaluation: Evaluation of Norwegian Development Cooperation with the Western Balkans

6.10 Study: Evaluation of Norwegian Business-related Assistance Uganda Case Study

5.10 Study: Evaluation of Norwegian Business-related Assistance Bangladesh Case Study

4.10 Study: Evaluation of Norwegian Business-related Assistance South Africa Case Study

3.10 Synthesis Main Report: Evaluation of Norwegian Business-related Assistance

2.10 Synthesis Study: Support to Legislatures

1.10 Evaluation: Evaluation of the Norwegian Centre for Democracy Support 2002–2009

2009

7.09 Evaluation: Evaluation of the Norwegian Programme for Development, Research and Education (NUFU) and of Norad's Programme for Master Studies (NOMA)

6.09 Evaluation: Evaluation of the Humanitarian Mine Action Activities of Norwegian People's Aid

5.09 Evaluation: Evaluation of Norwegian Support to Peacebuilding in Haiti 1998–2008

4.09 Evaluation: Evaluation of Norwegian Support to the Protection of Cultural Heritage

4.09 Study Report: Norwegian Environmental Action Plan

3.09 Evaluation: Evaluation of Norwegian Development Coopertation through Norwegian Non-Governmental Organisations in Northern Uganda (2003-2007)

3.09 Study Report: Evaluation of Norwegian Business-related Assistance Sri Lanka Case Study

2.09 Evaluation: Mid-Term Evaluation of the Joint Donor Team in Juba, Sudan

2.09 Study Report: A synthesis of Evaluations of Environment Assistance by Multilateral Organisations

1.09 Study Report: Global Aid Architecture and the Health Millenium Development Goals

1.09 Evaluation: Joint Evaluation of Nepal´s Education for All 2004-2009 Sector Programme

2008

6.08 Evaluation: Evaluation of Norwegian Development Cooperation in the Fisheries Sector

5.08 Evaluation: Evaluation of the Norwegian Research and Development Activities in Conflict Prevention and Peace-building

4.08 Evaluation: Evaluation of Norwegian HIV/AIDS Responses

3.08 Evaluation: Mid-term Evaluation the EEA Grants

2.08 Evaluation: Joint Evaluation of the Trust Fund for Enviromentally and Socially Sustainable Development (TFESSD)

2.08 Synthesis Study: Cash Transfers Contributing to Social Protection: A Synthesis of Evaluation Findings

2.08 Study: Anti- Corruption Approaches. A Literature Review

1.08 Evaluation: Evaluation of the Norwegian Emergency Preparedness System (NOREPS)

1.08 Study: The challenge of Assessing Aid Impact: A review of Norwegian Evaluation Practise

1.08 Synthesis Study: On Best Practise and Innovative Approaches to Capasity Development in Low Income African Countries

2007

5.07 Evaluation of the Development -Cooperation to Norwegion NGOs in Guatemala

4.07 Evaluation of Norwegian Development -Support to Zambia (1991 - 2005)

3.07 Evaluation of the Effects of the using M-621 Cargo Trucks in Humanitarian Transport Operations

2.07 Evaluation of Norwegian Power-related Assistance

2.07 Study Development Cooperation through Norwegian NGOs in South America

1.07 Evaluation of the Norwegian Petroleum-Related Assistance

1.07 Synteserapport: Humanitær innsats ved naturkatastrofer:En syntese av evalueringsfunn

1.07 Study: The Norwegian International Effort against Female Genital Mutilation

2006

2.06 Evaluation of Fredskorpset

1.06 Inter-Ministerial Cooperation. An Effective Model for Capacity Development?

1.06 Synthesis Report: Lessons from Evaluations of Women and Gender Equality in Development Cooperation

2005

5.05 Evaluation of the "Strategy for Women and Gender Equality in Development Cooperation (1997–2005)"

4.05 Evaluation of the Framework Agreement between the Government of Norway and the United Nations Environment Programme (UNEP)

3.05 Gender and Development – a review of evaluation report 1997–2004

2.05 – Evaluation: Women Can Do It – an evaluation of the WCDI programme in the Western Balkans

1.05 – Study: Study of the impact of the work of FORUT in Sri Lanka and Save the Children Norway in Ethiopia: Building Civil Society

1.05 – Evaluation: Evaluation of the Norad Fellowship Programme

2004

6.04 Study of the impact of the work of Save the Children Norway in Ethiopia: Building Civil Society

5.04 Study of the impact of the work of FORUT in Sri Lanka: Building CivilSociety

4.04 Evaluering av ordningen med støtte gjennom paraplyorganiasajoner.Eksemplifisert ved støtte til Norsk Misjons Bistandsnemda og Atlas-alliansen

3.04 Evaluation of CESAR´s activities in the Middle East Funded by Norway

2.04 Norwegian Peace-building policies: Lessons Learnt and Challenges Ahead

1.04 Towards Strategic Framework for Peace-building: Getting Their Act Togheter.Overview Report of the Joint Utstein Study of the Peace-building.

2003

3.03 Evaluering av Bistandstorgets Evalueringsnettverk

2.03 Evaluation of the Norwegian Education Trust Fund for Africain the World Bank

1.03 Evaluation of the Norwegian Investment Fund for Developing Countries (Norfund)

2002

4.02 Legal Aid Against the Odds Evaluation of the Civil Rights Project (CRP) of the Norwegian Refugee Council in former Yugoslavia

3.02 Evaluation of ACOPAMAn ILO program for "Cooperative and Organizational Support to Grassroots Initiatives" in Western Africa 1978 – 1999

3A.02 Évaluation du programme ACOPAMUn programme du BIT sur l'« Appui associatif et coopératif auxInitiatives de Développement à la Base » en Afrique del'Ouest de 1978 à 1999

2.02 Evaluation of the International Humanitarian Assistance of theNorwegian Red Cross

1.02 Evaluation of the Norwegian Resource Bank for Democracyand Human Rights (NORDEM)

2001

7.01 Reconciliation Among Young People in the Balkans An Evaluation of the Post Pessimist Network

6.01 Can democratisation prevent conflicts? Lessons from sub-Saharan Africa

5.01 Evaluation of Development Co-operation between Bangladesh and Norway, 1995–2000

4.01 The International Monetary Fund and the World Bank Cooperation on Poverty Reduction

3.01 Evaluation of the Public Support to the Norwegian NGOs Working in Nicaragua 1994–1999

3A.01 Evaluación del Apoyo Público a las ONGs Noruegas que Trabajan en Nicaragua 1994–1999

2.01 Economic Impacts on the Least Developed Countries of the Elimination of Import Tariffs on their Products

1.01 Evaluation of the Norwegian Human Rights Fund

2000

10.00 Taken for Granted? An Evaluation of Norway's Special Grant for the Environment

9.00 "Norwegians? Who needs Norwegians?" Explaining the Oslo Back Channel: Norway's Political Past in the Middle East

8.00 Evaluation of the Norwegian Mixed Credits Programme

7.00 Evaluation of the Norwegian Plan of Action for Nuclear Safety Priorities, Organisation, Implementation

6.00 Making Government Smaller and More Efficient. The Botswana Case

5.00 Evaluation of the NUFU programme

4.00 En kartlegging av erfaringer med norsk bistand gjennomfrivillige organisasjoner 1987–1999

3.00 The Project "Training for Peace in Southern Africa"

2.00 Norwegian Support to the Education Sector. Overview of Policies and Trends 1988–1998

1.00 Review of Norwegian Health-related Development Cooperation1988–1997

1999

10.99 Evaluation of AWEPA, The Association of European Parliamentarians for Africa, and AEI, The African European Institute

9.99 Evaluation of the United Nations Capital Development Fund (UNCDF)

8.99 Aid Coordination and Aid Effectiveness

7.99 Policies and Strategies for Poverty Reduction in Norwegian Development Aid