Global Evaluations and Research

Preregistered

The Global Evaluations and Research Programme

When do we learn the most from our evaluations?

NORWEGIAN CHURCH AID

actalliance

This evaluation forms part of Norwegian Church Aid's Global Evaluations and Research Programme.
The report was developed by Øivind Fjeld-Solberg, Quinn Coffey and Andrej Viotti. NCA©2022.

Layout: Steinar Zahl / Studio Zate

Cover photo: Håvard Bjelland/Norwegian Church Aid.
The right-holders depicted in the photos are not directly related to the results and/or findings presented in this report.

When do we learn the most from our evaluations?

Øivind Fjeld-Solberg, Quinn Coffey & Andrej Viotti

Methods, Evaluations and Learning, Norwegian Church Aid

Norwegian Church Aid's Global Evaluations and Research Programme

Global Evaluations and Research

Table of contents

NCA's Global Evaluations and Research Programme

"The main focus of NCA's Global Evaluations and Research Programme is to gather evidence and identify higher-level learning outcomes that can be broadly implemented in NCA's future programmatic work."

As a data-driven and results-based organization, high quality evaluations and research are seen as prerequisites for the constant refinement of Norwegian Church Aid's (NCA) work and how the organization evolves over time. This means that we utilize recognized, rigorous scientific methods when collecting data, when analysing data, and finally, when conclusions and recommendations are drawn or specified.

NCA's Global Evaluations and Research programme helps ensure that NCA's humanitarian and development assistance is as relevant, coherent, efficient, effective, sustainable, and impactful as possible. Therefore, the main focus of NCA's Global Evaluations and Research Programme is to gather evidence and identify higher-level learning outcomes that can be broadly implemented in NCA's future programmatic work.

The following criteria informs the topics selected for the programme:

- The topic is relevant, i.e., in concord with NCA's 2020-2030 Programme Framework.
- The topic is global, i.e., covering work in at least two countries.
- The topic is applicable, i.e., findings can be broadly implemented in the organization.
- The topic covers an existing knowledge gap.

As of 2021, the Global Evaluations and Research Programme follows NCA's "PATH-framework for evaluation and research". PATH is an acronym for Preregistered, Accountable, Transformative and Honest evaluations and research. By following the PATH-framework, we ensure that all global evaluations and research projects are as transparent, objective and scientifically rigorous as possible. In practice this means that the programme focuses on fewer, more in-depth evaluations and research projects that utilize best practice scientific methods and dissemination strategies that ensure high levels of implementation and reintegration of specific and actionable evidence-based recommendations in the organization.

As stated in the PATH-framework, all of NCA's global evaluations and research projects are preregistered in order to ensure full transparency. Preregistration is the practice of registering a scientific study/evaluation *before* it is conducted. Preregistration of NCA's global evaluations and research serves to enhance trust in NCA's evaluative work. Specifically, preregistration helps readers distinguish between hypothesis-generating (exploratory) and hypothesis-testing (confirmatory) aspects of our evaluations and research. By defining and registering key questions, hypotheses, methods, and an analysis plan before we observe the outcomes, we help prevent bias, reduce data dredging, and avoid hypothesizing after the results are known.

As stated by Norad – **our strength lies in facts or evidence**. In order to be truly effective as an organization, we therefore need to craft programmes and interventions that are based on evidence – evidence that tells us what benefit right-holders the most. The Global Evaluations and Research Programme aims to help NCA reach that goal.

Executive Summary and Recommendations

In this evaluation we asked the question; "When do we learn the most from our evaluations?". The goal was to identify areas in need of refinement and identify best practice evaluative processes. A cross-sectional mixed method design was utilized, including an online survey and key informant interviews. Based on a synthesis of the results, the answer to the evaluation question seems to lie in the following five main outcomes.

1. We learn the most when the quality of the management response is high
2. We learn the most when specificity and feasibility of recommendations and management responses are high
3. We learn the most when the context is considered in evaluations
4. We learn the most when the timing of evaluations is right
5. We learn the most when organizational barriers to learning are few

In response to these findings, the following specific recommendations are outlined in the report.

Improve management response letter development and follow-up

In order to improve management response letter development and follow-up, we recommend making changes in the following order:

Policy development

Revise NCA's Evaluation and Research policy and chapter 10.3 in the Operations manual. Changes should reflect the importance of the management response letter with revised routines for its use. The management response letter template is also in need of revision. The use of the new template should be made mandatory and the letter itself should include more mandatory elements and detailed instructions on how to complete it. The transformative part of the PATH-framework should also include an added emphasis on the use of management response letters in the pursuit of our evaluations' true transformative potential.

Adjustment in routines

There is a need to build new routines for management response follow-up according to the deadlines set in the management response letter. This should be done by utilizing the full potential of PIMS, building a separate evaluation workflow that correspond to mandatory elements in the evaluation process. The workflow, with prespecified tasks, should be utilized to issue reminders to identified staff, including instructions for follow-up of specific tasks and deliverables.

Capacity building

There is a need to build capacity in the completion of management response letters and how to translate evaluation findings into specific, actionable tasks that can be implemented in a prioritized order, including clear deadlines and a division of labour and responsibility. The importance of NCA's management response letter and its use should also be made known to everyone through a new launch and roll out.

Focus on context and contextualization

In order to increase contextual understanding and the contextualization of evaluations, we recommend making changes in the following order:

Policy development

Evaluations must, to a larger degree, take local context into consideration and field visits should be strongly recommended, including direct contact with key stakeholders and right-holders through user-involvement in key phases of the evaluation process. Include a paragraph on context and contextualisation in the revised Operations manual and Evaluation and Research Policy. The accountable part of the PATH-framework should also include an added emphasis on the importance of contextualization and user involvement when establishing accountability.

Adjustment in routines

A shift in routines towards more internally led evaluations should be implemented. Internally led evaluations will ensure better understanding of the local context and secure contextualization of evaluation recommendations. When external consultants are used, we need to ensure that they are familiar with the local context and that they conduct field visits when possible. NCA's Evaluation report and recommendations template is in need of revision. Focus should be on learning, and contextual considerations should be included in learning outputs to a larger degree.

Capacity building

There is a need to build capacity on the importance of contextual consideration in evaluations. Capacity building in quality assurance of bids and evaluation inception reports should also be prioritized. When giving bids a technical score, "tenderers relevant experience in the field of assignment" and "tenderers experience in the region/country e.g., knowledge of local language, culture, administrative system, government etc." should be given added weight in order to reflect the emphasis on context and contextual knowledge.

Improve timing and commitment to learning in evaluations

In order to improve the timing of evaluations, we recommend making changes in the following order:

Policy development

Revise NCA's Evaluation and Research policy and Operations Manual, with an increased emphasis on the timing of evaluations and the utility of mid-term evaluations in learning exercises. A reduction in the number of evaluations conducted each year and the cancelation of evaluations that focus on short, one-year projects where there are no donor requirements for an evaluation, should also be discussed.

Adjustment in routines

NCA's midline evaluations and mid-term reviews should be prioritized when learning is to be extracted. All Terms of References (ToRs) for midline evaluations and reviews should include specific paragraphs where a clear intention for learning is expressed. How knowledge and learning outcomes from the evaluation will be implemented and reintegrated in the programme being evaluated and in the wider organization should also be stated.

Capacity building

Midline evaluation and mid-term reviews should to a larger degree be conducted internally in order to ensure high level of staff involvement and knowledge harvesting. It is therefore a need to build capacity in internal evaluation management and learning outcome dissemination through internal workshops and seminars. Management must also make sure there is time/room to reflect on, implement and disseminate findings from evaluations in staff schedules, incorporating e.g., reflective sessions on learning as an individual PDR goal, while actively monitoring and adjusting individual staff workload.

Reduce barriers and enable a culture for learning and knowledge reintegration

In order to decrease the number of barriers to learning from evaluations, we recommend making changes in the following order:

Policy development

NCA needs to build a stronger evaluation and learning culture and develop a shared understanding of the purpose of evaluations in meeting our learning goals. The PATH-framework's "transformative" section should be revised to reflect this and help build a culture for learning, enhancing staff understanding of the importance of learning, evidence, and the scientific approach to effective development and humanitarian aid.

Adjustment in routines

A culture for learning can also be enabled by systematically and routinely feeding lessons learnt back into organizational planning and decision-making, e.g., via regular updates from GERA in DIP-meetings. Webinars and a wide circulation of the Evaluation and Research Bulletin as an "easy to access" format would also strengthen such a culture. All NCA's evaluations should be made widely available and searchable for everyone through a dedicated Learning site containing an Evaluations library. This will make it easier to extract learning from evaluations and get an overview of the knowledge previously gathered in a specific thematic area or country.

Capacity building

There is a need to develop staff's capacity to produce credible, scientifically rigorous evaluations internally and/or in collaboration with external consultants. Capacity building in the use of a new PIMS evaluation workflow, quality assurance of inception reports, reports and recommendations should also be prioritized. This will reduce barriers to learning and enable staff to extract learning from evaluations and the evaluation process itself more fully. Staff involved in evaluation and/or programme development should routinely extract learning from previously conducted evaluations and documented learning processes.

Abbreviations and Keywords

NCA: Norwegian Church Aid
HO: NCA's Head Office in Oslo
CO: Country Office
MEL: The Methods, Evaluations and Learning team in Oslo
ToR: Terms of Reference

Intervention: When the term "intervention" is used, it generally refers to the subject of the evaluation. Here, interventions encompass all the different types of development and humanitarian efforts that may be evaluated, such as a project, programme, policy, strategy, thematic area, technical assistance, policy advice, an institution, financing mechanism, instrument, or other activity.

Introduction

Over the years, the prevalence and importance of evaluations has increased. Stakeholders, whether they are donors or partners, want to be assured that their investments (e.g., money, resources, time) are used efficiently, make a difference and uphold ethical principles. As the importance of evaluations has become more evident, evaluative work has also become more professionalized as a practice.

Still, there is a growing concern that evaluations, to a large degree, have become mere "measures of reassurance", i.e., ways of documenting results and progress for donors and other stakeholders (Cracknell, 2000). Indeed, evaluations play a vital role in such reporting schemes, but high-quality evaluations of programmes and/or interventions should not only provide information about e.g., efficiency, effectiveness and/or target achievements – they should also consider the "why" and "how" in this equation. We need to ask why desired outcomes were achieved or not achieved, and how we can improve interventions or programmes going forward. If findings from evaluations are not being fully translated into learning outcomes, by identifying the active ingredients of change and potential for improvement, valuable learning opportunities can be missed.

According to NCA's new strategic priorities, we want to strengthen our ability to obtain relevant results, analyse our achievements and identify areas for improvement in order to learn. This can only be achieved if we move beyond accountability and reassurance reporting, and utilize the full potential of our evaluative work, reintegrating what we have learnt in the wider organization. In turn, this will increase the impact of donor investment and the resources spent in the field.

Purpose of the evaluation, audience, and use

With the above challenges in mind, the purpose of this global evaluation was to establish a "baseline" or "snapshot" of evaluation quality and knowledge reintegration in NCA, asking the following question:

When do we learn the most from our evaluations?

At the same time, we wanted to explore the "why" and "how" in the equation. Why are evaluation findings not fully reintegrated in the organization? How can we utilize the full potential of our evaluations? What is hampering the reintegration of knowledge? And when we get it right, what brings this about? The overall aim was to identify when we learn the most from our evaluations and establish a best practice framework for NCA's future evaluative work going forward. Our target audience was all HO and CO staff involved in either the execution of evaluations or the use of evaluation learning outcomes in programme development and/or refinement.

Evaluation criteria and questions

The purpose of the present evaluation was to establish a baseline of evaluation quality and knowledge reintegration in NCA. The aim was to improve learning. Our main hypothesis was that *after an evaluation is conducted, high levels of learning outcome implementation, learning and knowledge reintegration in the organization is predicted by specific, actionable and feasible recommendations and management response letters that include specific, high quality action points, a specified timeline, and a clear division of labour/ responsibility.* In other words, evaluations that are a priori planned and designed well, with the specific intent of learning, will produce the most actionable recommendations and learning outcomes (See preregistration form for details, Appendix IV).

The following DAC-criteria were used as sub-questions or lenses of focus:

Coherence:

Are evaluations in general planned, designed and executed with the specific goal/intent of producing actionable recommendations that can be implemented in order to improve our work?

Effectiveness:

Are evaluation recommendations and management responses, when presented, implemented in order to improve programming, project designs and implementation strategies? If not, why? What are the barriers? How can this be improved?

Impact:

When evaluations are used in a constructive way, what brings this about? Who is implementing recommendations and how is this done? How can we enhance the impact and reintegration of lessons learned going forward? What needs to improve? What are the obstacles?

Method overview

The geographical scope of the evaluation was global and included a retrospective desk review of evaluations, a cross-sectional survey, and key informant interviews. The sample frame consisted of all evaluations conducted in 2018 and 2019 and their related management response letters. In total 25 evaluations and their respective recommendations and management response letters were identified. A more detailed description of the methods used is annexed (Appendix I).

Online survey and quality ratings

A survey consisting of both quantitative and qualitative measures concerning evaluation quality and learning

was distributed to HO and CO staff who were involved in the 25 selected evaluations (See Appendix II for the full questionnaire). All included evaluations were also rated in relation to the quality of the recommendations presented and the corresponding management response letter by NCA's Methods, Evaluations and Learning team (MEL).

Key informant interviews

Ten of the 25 evaluations were selected for in-depth, semi-structured interviews with key informants. Evaluations were selected based on MEL-ratings (5 "high" and 5 "low" rated evaluations). Key informants included CO staff directly involved in planning of evaluations and/or the development of management responses.

The interviews lasted approximately 1 hour and covered four main themes. The themes covered were "Overall quality of the evaluation", "the Evaluation Process", "the Evaluation Response", and general questions regarding "the usefulness of the evaluations and NCA's approach to evaluations". (See Appendix III for interview guide).

Results overview

In this section we briefly describe the quantitative and qualitative results of the evaluation separately. A more detailed result section is annexed (Appendix I)

Quantitative results

In total, we received 26 survey responses related to 20 evaluations. 5 evaluations did not receive any responses and were therefore excluded from the results. The MEL-team rated all evaluations included.

Overall quality rating from the MEL-team

All included evaluations were rated in relation to the quality of the recommendations presented (specificity and feasibility) and the corresponding management response letter (specificity, division of labour and deadlines) by three independent raters in MEL. Results revealed that raters from the MEL-team overall rated the quality of recommendations and the management response letters slightly lower than CO staff. Moreover, MEL-raters reported that most recommendations and management responses were hard to rate due to their brevity and general lack of detail and use of generic language.

Survey results

On average, survey informants rated the overall quality of the evaluations (methods, report, summary, management response letter) from as "good" to "very good". Responses varied from "poor" to "very good". Similar responses were found in relation to the average perceived value of recommendations and action points presented in the report summaries and management response letters. Again, responses varied to some degree, but on average the value of recommendations and management responses were rated to be of "high" to "very high value" (see table 1 and 2 in Appendix I for details).

When informants were asked to what degree the recommendations/action points in the management response letter were implemented in future programming, the average score given indicated that recommendations/action points were only implemented to a moderate degree. Responses ranged from "not at all" to "a very high degree".

Similar responses were given when rating learning for future programming and reintegration of knowledge gained from the evaluation in the organisation. Again, informants reported that the recommendations/action points presented in the management response letter only produced learning for future programming to a moderate degree and that reintegration of knowledge was done to a moderate degree. Responses ranged from "a low degree" to "a very high degree" (see table 3 in the Appendix I for details).

Finally, when asked to what degree informants felt that the evaluation produced actionable and relevant learning outcomes that could be integrated in the future and how satisfied they were with the evaluation process and its outcomes, the average score given indicated that actionable and relevant learning outcomes were produced to a "high degree" and that informants in general were "satisfied" with the evaluation and its outcomes (see table 4 in Appendix for details).

Associations between quality, learning and knowledge reintegration

As described in the preregistration form (see Appendix IV), our main hypothesis was that learning outcomes and knowledge reintegration would be associated with the overall quality and value of summary recommendations and management response letters, and that the most impactful summaries and response letters would be rooted in evaluations with high quality methods (design, data collection and analysis) and/or high-quality reports. When analysing associations between these factors, we found some evidence supporting our main hypothesis, but a somewhat more complex picture than expected emerged, as shown in figure 1.

As shown in figure 1, the variable "overall quality of the management response" was positively associated with "the degree to which recommendations/action points presented in the management response letter produced learning for future programming" and "the degree to which the knowledge/learning gained from the evaluation was reintegrated in other areas of the organization" as predicted. However, the "overall quality of the evaluation summary and recommendations" were not associated with "the

degree to which recommendations/action points presented in the management response letter produced learning for future programming" or "the degree to which the knowledge/learning gained from the evaluation was reintegrated in other areas of the organization". Finally, the quality of recommendations and management responses were not significantly associated with the "degree to which recommendations/action points in the management response letter were implemented/utilized in future programming" (See Appendix I for full details).

Key informant interviews

In this section we present key results from the qualitative part of the evaluation, including key informant interviews and responses to the three "qualitative" items from the survey where informants could write responses to open ended questions. The interviews explored several quality factors that contributed to the implementation of learning from evaluations, as well as knowledge sharing and retention. The most central themes that emerged from the interviews are discussed here.

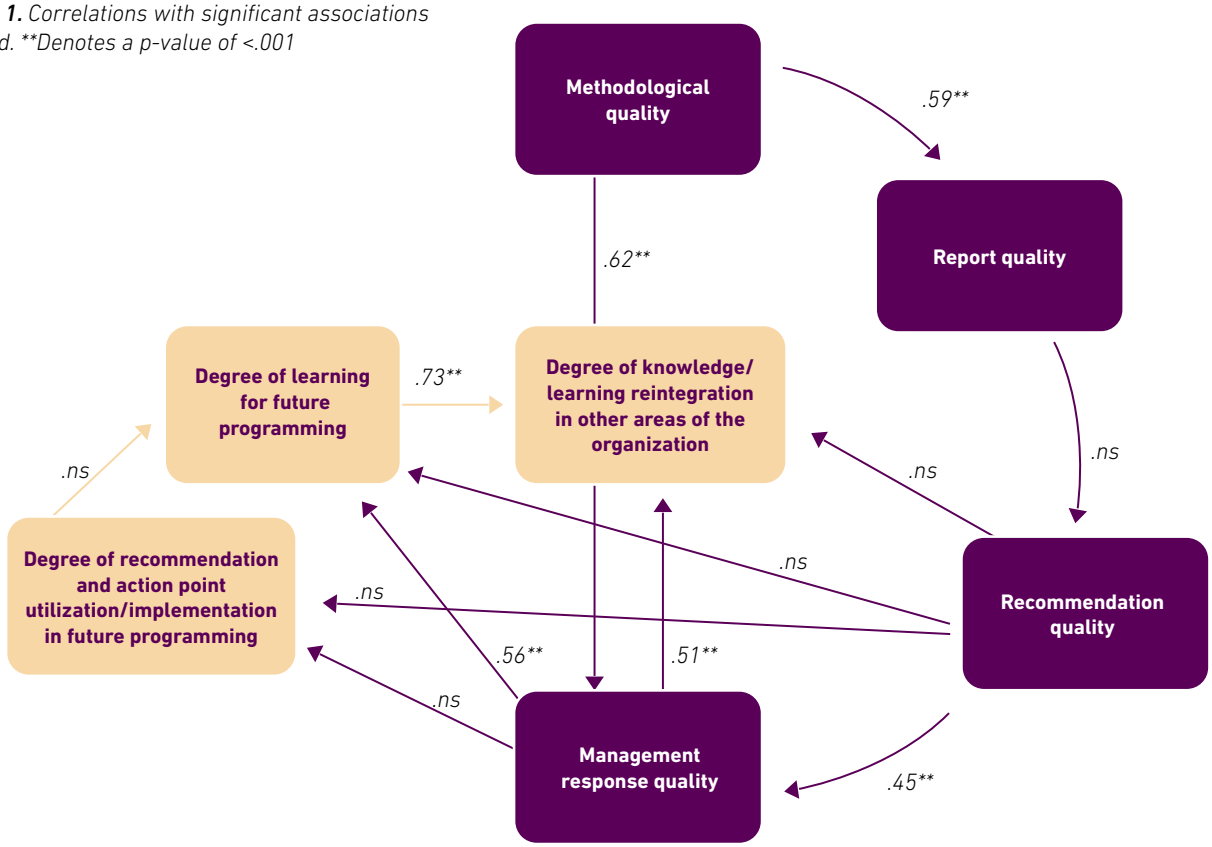
General quality and usefulness of recommendations and management response letters

As shown in the quantitative section, overall quality of recommendations and management responses were in general viewed as being of good quality. This finding was supported in the qualitative interviews, where recommendations and management responses were generally viewed in a positive to neutral light. Still, in some instances, informants were not aware of the requirement for a management response letter or what template to use for this type of response.

> *"(We do) not usually use a management response letter. We work together. Management response is in the minutes from the meeting, becomes the action plan. We do not have a full overview of the tools and templates."* [1]

Another factor contributing to the usefulness of evaluations was the specificity and feasibility of evaluation recommendations. Recommendations that were seen as overly broad, were typically disregarded. In some cases, recommendations were seen as either beyond the scope of the project or financially or logistically impossible. In general, there

Figure 1. *Correlations with significant associations marked. **Denotes a p-value of <.001*



1) Interview 11

was an association between the perceived quality of the evaluation and the specificity/feasibility of the evaluation recommendations.

> *"Recommendations are useful when they are specific -- especially those that don't need a huge resource investment."* [2]

Additionally, recommendations that included some practical guidance on how they should be implemented, were viewed as useful.

> *"Recommendations that outline the practical aspects of the evaluation and make a strategy for how to improve the programme or support to partners are useful. This is a gap that needs to be improved. It is influenced by the quality of the evaluation."* [3]

Contextuality

The degree to which an evaluation was seen as contextually relevant played a strong role in whether or not it was viewed as useful. Here contextuality refers both to the consultants having a clear understanding of the local context and being familiar with the international development field. In fact, in order for the evaluation to be contextually relevant, the need for consultants to travel to the field was viewed as a necessity across several interviews.

> *"There were, however, some areas that could be improved -- primarily that the consultant wasn't able to travel to meet with relevant stakeholders in person. This would have given the evaluator a better understanding of the context in which the project was being implemented. The evaluations that are conducted totally remotely are not that relevant compared to field visits"* [4]

In the above instance, the respondent found the report to be overly broad and felt that it could be improved, and that the recommendations would be more specific, if the consultant had spoken directly to beneficiaries and partners.

In another instance, the respondent found the evaluation to be useful, but the report itself was not written within an international development context, which impacted its utility.

> *"The evaluation methodology was good, but the consultant, who was from a Norwegian university, took more of an academic approach which affected how relevant the recommendations were. They lacked a development perspective."* [5]

Data collection was also seen as impacting the utility of evaluations. Several factors seemed to compromise data quality, e.g., resource constraints that impacted the scale and feasibility of data collection, limited access to beneficiaries due to conflict, non-representative samples and the relevance of the data that was collected.

> *"On the negative side, there were some challenges with data collection because in some areas there were either no informants or they were difficult to access because of the context."* [6]

Challenging assumptions and ways of working

According to informants, evaluations that challenged their assumptions or NCA's programmatic methodologies were seen as highly relevant and useful. A majority of informants appreciated the role of evaluations as a type of third-party audit of NCA's work.

> *"Evaluations [are useful] when they come up with something that challenges what we have been doing, and what we considered that we did well. Evaluations might say actually that is not true."* [7]

Evaluations also seem to play a role in highlighting gaps and challenging the status quo. This was particularly true of midline evaluations (discussed later). Evaluations that challenged assumptions also played a role in inspiring new directions for programming or giving some weight to strategic decision making.

> *"Evaluations [are useful] when they are an inspiration for new proposals and inspire change in programmatic decisions. They help us realize that we need to change something."* [8]

Balance between positive and negative

Evaluations that managed to strike a balance between highlighting positive results, and giving constructive feedback were also perceived as more useful – particularly when addressing recommendations.

> *"When recommendations are only negative.... we need some positives as well. It doesn't motive us to implement them...they must not demoralize."* [9]

Frequency and timing

Finally, the frequency and timing of evaluations played a significant role in their perceived usefulness. In general, midline evaluations were viewed as more useful than endline evaluations because staff felt that they could more easily make use of recommendations from midline evaluations while programmes were ongoing.

2) Interview 2 3) Interview 9 4) Interview 1 5) Interview 9 6) Interview 2 7) Interview 7 8) Interview 3 9) Interview 11

"The best learning comes from midterm evaluations. Then we can address problems while the project is ongoing" [10]

Along similar lines, evaluations of short term (e.g., one year or less) projects were viewed in a very negative light.

"Evaluation of short projects is a waste of resources. They do not go into depth." [11]

"It is most useful to have mid and endline evaluations. The project needs to have time to develop and reach their goals [before being evaluated]." [12]

Still, regular evaluations were generally seen in a positive light, as they acted as checks on long term programmes. Endline evaluations were seen as useful mainly when a similar project was to continue in the future.

"Evaluations should be done more often (depending on needs) because this is the mirror that you look at yourself otherwise you don't know if you're doing it right or not. Especially when done externally. This is like our audit. They maybe should be done every 2 years. To make changes before it's too late." [13]

"Most useful when it comes to longer term projects greater than a year. Midline is also useful if it is on time, with immediate effect on ongoing projects. On the other hand, the end line evaluations are useful because it can have a good impact on future projects during project design." [14]

Barriers to learning

Beyond quality and usefulness acting as a predictor of learning/implementation, this evaluation also found that other factors acted as barriers to learning. In general, informants saw the value of extracting and reintegrating learning from evaluations in their work. Knowledge sharing beyond NCA was also valued. However, several barriers to learning and knowledge sharing were identified, and several informants suggested that NCA lacked routines for learning capture and knowledge sharing.

"We need to better the follow up of action plan/ monitor the change/learning. No established routines for follow up of action plan and monitoring of implementation... Not very good routines for sharing, little sharing. Something we (CO) need to improve." [15]

In other cases, informants connected knowledge sharing/learning from evaluations to the preservation of institutional memory in NCA, suggesting that new staff were not always exposed to, or had access to older evaluations which might help in their work.

"[Evaluations are] always useful somehow, but when they are not taking account for future activities, new staff cannot understand the lessons learned. Need to be communicated to new staff." [16]

In this case, informants suggested that learning had taken place, but again learning was not actively shared going forward – partly because of a lack of routines, but also because of NCA's own information management systems.

The "why?" questions

In the quantitative survey, 3 questions of a more qualitative form were included in order to directly probe informants for information regarding why recommendations, action plans and learning outcomes were not utilized or reintegrated in the wider organization. A selection of the answers given are outlined below.

If recommendations/action points were not implemented/utilized to a high degree, why was this so?

When investigating the answers given to this question, we found a large variance in "reasons" for not implementing recommendations. Still, some themes emerged related to resources, relevance, contextualization, and changes in programming.

"Probably because of limited inclination of the implementing officers to learning and of the limited supervision capacity of the CO managers / SMT."

"Recommendations are rarely followed. It focuses more on the execution and follow-up of projects to achieve annual results (...)"

"Recommendations were not contextualized (...)"

"The recommendations were generic and not very specific to the context."

"Change of Strategy affected the implementation of certain recommendations."

The timing of evaluations also played a role in the degree to which recommendations were implemented.

"The Mid-term review was conducted in the beginning of 2019 in the end of the strategy meaning that there was no time to implement the recommendations (...)"

10) Interview 10 11) Interview 4 12) Interview 5 13) Interview 9 14) Interview 1 15) Interview 3 16) Interview 5

"Despite some good adoption on the implementation of recommendations, there is a need to create a space to reflect on the implementation of proposed recommendations."

If recommendations/action points did not produce learning to a high degree, why was this so?

When investigating the answers given to this question, lack of resources and capacity were stated as the main reasons for low levels of learning.

"Limited capacity of the managers / SMT to mentor and coach officers."

"The implementation of the recommendations/action points couldn't be achieved much due to shortage of resources (funds) to improve capacities and infra-structures (...)"

If the knowledge/learning gained were not reintegrated to a high degree, why was this so?

Finally, when investigating the answers given to this question, we again found a large variance in "reasons" for not reintegrating knowledge and learning. Answers related to relevance, feasibility, and access to information were common.

"The project is very contextual (...), hence our other organizational areas cannot adopt that learning extensively."

"We are still utilizing the learning, but again the change in the thematic focus affected the utilization of all the learning points."

"It depends on the practicality and feasibility of the recommendations. If the recommendations are valid, they should be practically implemented and given importance in future programme designing."

"Access to information with respect to earlier evaluations is not available to new staff. It is recommended that proper orientation may be provided in the form of access to earlier evaluation studies so that new staff could learn and incorporate the learnings."

Discussion of results and findings

> ***"The overall methodological or scientific level of evaluations might be one of the main foundational underpinnings of learning and knowledge reintegration in the organization."***

In NCA, learning is the primary purpose of evaluations and research. In order to learn and evolve as an organization, we need to understand what we have achieved in the past, where major challenges still lie ahead, and finally how these challenges can be addressed. In the present global evaluation, we therefore asked the question; "When do we learn the most from our evaluations?". The goal was to identify areas in need of refinement and identify best practice evaluative processes. Based on a synthesis of the results from both the interviews and the survey, the answer to this question seems to lie in the following five main outcomes.

1. We learn the most when the quality of the management response is high

Our main hypothesis for the present evaluation was that high levels of learning outcome implementation and knowledge reintegration in the organization would be predicted by specific, actionable and feasible recommendations and management response letters that included specific, high quality action points, a specified timeline, and a clear division of labour and responsibility. As described in the results section of this report, we found evidence supporting this hypothesis, but findings paint a less straight-forward relationship between the factors.

Indeed, when analysing the association between quality predictors and learning outcome dimensions, the variable "overall quality of the management response", was positively associated with "the degree to which recommendations/action points presented in the management response letter produced learning for future programming" and "the degree to which the knowledge/learning gained from the evaluation was reintegrated in other areas of the organization". However, the "overall quality of the evaluation summary and recommendations" were not associated with "the degree to which recommendations/action points presented in the management response letter produced learning for future programming" or "the degree to which the knowledge/learning gained from the evaluation was reintegrated in other areas of the organization".

In sum, this finding seem to identify a higher degree of adherence to the action points presented in high quality, high value management response letters, compared to recommendations presented in the evaluation report summary. This might possibly reflect that following up on management response letters are mandatory, while recommendations that are not prioritized in management response letters can be postponed or disregarded to a larger degree. The findings might also reflect a dearth of high quality, useful, feasible and relevant recommendations in evaluation reports, as suggested in several of the interviews. This might result in a prioritization of the recommendations identified in management response letters only.

Interestingly, neither the quality of recommendations nor management responses were significantly associated with the "degree to which recommendations/action points in the management response letter were implemented/utilized in future programming". This might reflect the possibility that factors other than recommendations and management response letters seem to influence the degree of implementation of recommendations/action points in future programming. When we asked directly why recommendations were not implemented to a high degree in the quantitative survey, responses from

informants added corroborative evidence to this notion. In their answers we found, as mentioned, a large variance in "reasons" for not implementing recommendations and action points that were not directly related to the recommendations or the management response letters. Instead, several instances of low-level implementation were reported to be caused by *limited capacity of managers, change of strategy, the timing of the evaluations, no time to implement or no time to reflect on recommendations.*

Finally, one of the strongest associations in the quantitative analyses was found between overall methodological quality of the evaluations and management response letter quality. This means that the design, data collection and statistical analysis, i.e., the scientific quality of the evaluations, were associated with specific, high quality action points in the management responses. Considering that high quality management responses seem to be so important for learning, this is an important finding. Moreover, methodological quality was also associated with the overall quality of the evaluation reports. In essence, this seems to indicate that the overall methodological or scientific level of evaluations might be one of the main foundational underpinnings of learning and knowledge reintegration in the organization.

2. We learn the most when specificity and feasibility is high

In the interviews, it became clear that we seemed to learn more from our evaluations when the specificity and feasibility of recommendations in reports and action points in management response letters were high. This means that when action plans were formulated in a specific language, with practical guidance on how they could be implemented, they were perceived as more useful and easier to integrate in future programmatic work. Action plans that were feasible, i.e., possible to implement given time, financial and contextual constraints, were also perceived as having more utility.

Again, in light of the above-mentioned lack of association between recommendations, action plans and the implementation of these potential improvements in future programming, this finding might add another piece to the puzzle. If recommendations/action points are not specific enough and/or feasible, they might be difficult to interpret and/or hard to implement and extract learning from. This notion is also supported by the MEL-team's moderate ratings of recommendations and management response letters, and the general feedback that most recommendations and management responses were hard to rate due to their brevity and/or general lack of detail and the use of generic language.

3. We learn the most when the context is considered

We seemed to learn the most from evaluations when evaluations were contextually relevant, i.e., the consultants conducting the evaluations had a good understanding of international development assistance and spent some time in the field with direct contact with key stakeholder and right-holders. In fact, when a deep knowledge of the context and the local right-holders situation was embedded in the evaluation recommendations, the evaluation was perceived to have a higher degree of utility.

This finding might be linked to or reflect the previously mentioned lack of association between recommendations, action plans and the implementation of these changes or potential improvements in future programming. If recommendations/action points cannot be easily translated into the local context or does not take the context into consideration, recommendations and action points may be perceived as not relevant or hard to implement due to a lack of local fit. Local ownership is important here. Recommendations or action points for which there is a strong sense of local ownership are much more likely to implemented and better sustained than recommendations/action points which local staff/partners regard as recommended or imposed by external consultants without taking the local context into consideration.

4. We learn the most when the timing is right

On a positive note, informants in general saw the value of extracting and reintegrating learning from evaluations in their work. Knowledge sharing was also valued. However, results showed that several factors, like timing, were hampering learning, knowledge-sharing and reintegration in the organization.

In general, informants stated that midline evaluations were more useful since learning outcomes, specified as recommendations and corresponding action plans, could be used to inform, change and/or refine ongoing programmes by challenging assumptions and our ways of working. On the other hand, evaluations of shorter projects, e.g., evaluations of one-year projects, were seen as less useful, and in some instanced not useful at all/a waste of resources. This finding could be used to inform NCA's evaluation and research policy by adjusting the requirements related to the timing and frequency of evaluations.

5. We learn the most when organizational barriers to learning are few

In the interviews several informants suggested that NCA lacked routines for learning capture and knowledge

sharing. Informants also reported that there was often little or no time to implement, nor time to reflect on the recommendations presented. Finally, informants also connected the lack of knowledge sharing/learning from evaluations to a fragmented institutional memory in NCA, suggesting that new staff were not always exposed to, or had access to older evaluations that might help them in their work. In fact, the insufficient use of evaluations was also linked to high staff turnover and high workloads that might limit motivation and available resources for evaluation follow-up. Removing these barriers would increase the learning potential for staff and free up time to reflect on, understand and implement the recommendations given.

Limitations

A proper discussion of the limitations of the present evaluation is highly warranted in order to understand the scientific limitations of the presented findings. Without this level of clarity, it is impossible to interpret the validity of the scientific work or assess the credibility of the conclusions. So, although the present evaluation has several strengths, some key limitations also need mentioning.

Cross-sectional design

The design of the present evaluation is cross-sectional, meaning that all data was collected at one time point. The data collection is therefore relatively quick and inexpensive to conduct, but the design has some important limitations. The primary limitation of cross-sectional evaluation is that the temporal link between the measured variables and outcome cannot be determined because both are examined at the same time. In other words, no causal inferences can be drawn. In the present evaluations we therefore write about associations and relationships between variables and outcomes. In the places where causal directions are inferred, these inferences are based on a priori assumptions and hypotheses presented in the preregistration form (Appendix IV).

Timing

The present evaluation focused on evaluations conducted in 2018 and 2019, not evaluations conducted in 2020 and 2021. This was done in order to exclude evaluations that could have been impacted by the COVID-19 pandemic. The reasoning behind this was that the evaluations conducted in these years would be of lower or non-representative quality. In turn, this would bias the findings and limit the usefulness of the present evaluation. However, this also means that we cannot determine whether the associations and findings found in this report fully reflect the current state of evaluation quality and evaluation processes in NCA. Moreover, since evaluations included in this evaluation were conducted some time ago, it might also have been difficult for the survey informants and key informants to retrospectively recall in detail the content and quality of the evaluations in question, potentially adding a level of uncertainly to the results.

Data collection and sample size

In the present evaluation, we utilized a non-random sampling method, including all evaluations conducted in 2018 and 2019. Based on ratings (High/Low) from the MEL-team, 10 evaluations were then selected for key informant interviews. Moreover, survey data was only collected from 26 CO staff members. This non-random, limited scope, small number of interviews and survey informants might have influenced the representativeness of the findings presented. In other words, conclusions drawn in this report might have changed somewhat if we used a randomized design, and a larger number of informants participated in the survey and/or we had the resources to interview more staff.

Subjectivity and bias

The present evaluation used a self-report questionnaire and key informant interviews as a means of data collection. This means that all results are subjective in nature and not based on actual numbers from monitoring or annual reporting. In other words, the results presented only reflect the individual, subjective opinions of the survey respondents and key informants without any verification of these opinions in documented change/impact in e.g., country office reports or evaluation reports. This could potentially have resulted in an acquiescence bias.

Acquiescence bias or agreement bias is the common tendency for survey informants to agree with research statements, presenting desirable opinions/answers without the answer being a true reflection of their position or thoughts on the topic. In other words, the findings presented in this report can be somewhat positively biased. Considering the small number of informants, this limitation should be considered when recommendations are made on the basis of the presented findings.

Adherence to preregistration

One of the major strengths of this evaluation is its preregistration. As far as we know, this is the first evaluation in Norwegian aid operations history to date that has been officially preregistered. Preregistration is the practice of registering a scientific study/evaluation before it is conducted. This help readers distinguish between hypothesis-generating (exploratory) and hypothesis-testing (confirmatory) aspects of the evaluation. By defining and registering key questions, hypotheses, methods, and an analysis plan *before* we observe the outcomes, we also help prevent

"The use of external consultants might not be the panacea for bias and selective reporting donors are looking for. In fact, in some of the interviews informants reported that consultants were the cause of such bias, rather than the factor protecting against it"

bias, reduce data dredging, and avoid hypothesizing after the results are known. Still, some limitations must be mentioned in terms of the adherence to the preregistered analytical plan.

Considering that we found little variance in scores for the selected evaluations in terms of overall evaluation quality and management response letter quality, it was not possible to fully discern high-quality evaluation from low-quality evaluations and/or low-quality management responses from high-quality management responses, and grouping them as such. Analyses using analyses of variance (ANOVAs) were therefore dropped, keeping only the correlation matrix as our main analysis. ANOVAs are best suited when two or more populations/samples are compared, focusing on a potential relationship between the independent and the dependent variables. Since we could not fully identify the main independent variable (high quality vs. low quality evaluations) in this study, as specified in the preregistration from, we decided to look at correlations or associations only.

The items concerning the *value* of recommendations and management response letters were also excluded from the evaluation report. When running correlations for these items in relation to learning and knowledge integration, the results were equal to the pattern of correlations found for the *quality* of recommendations and management response letters. Since we originally did not include a clear definition of what was meant by the term "value" in the context of this evaluation, the added value of including these items in the analyses was deemed low considering that neither the informants, nor the readers of this report could fully understand the subtle difference between "quality" and "value" these items intended to capture.

Reflections

An evaluation that limits its view to inputs and outputs, or that only instrumentally documents a process, ignoring the complexity and context of our work, will not be able to produce findings that can influence subsequent behaviour and create learning in the organization. In this section of the report, we therefore venture slightly beyond the data gathered in this evaluation to reflect on a few additional themes related to the context of our work.

We learn the most from our evaluations when scientific quality/rigorousness is high?

Interestingly, scientific quality/rigorousness was not mentioned in any of the interviews. In other words, when we asked, "when are evaluations the least useful?", the theme of methodological quality was mentioned in passing, but none of the informants answered, "when evaluations have low scientific quality or lack of scientific rigorousness". Nor did they mention this as a factor that might impede usefulness or learning outcomes. This is more an observation than an explicit, data-driven finding. Still, this might allude to an underlying insensitivity to the importance of scientific evidence and the role that scientific rigorousness plays in high quality evaluations. The importance of reflecting upon this is also underlined in the findings that the methodological quality of the evaluations was associated with both the overall quality of the evaluation reports and specific, high quality action points in the management responses. In our view, this must be followed up in order to create a sense of urgency for an improvement in both scientific quality and organizational understanding of the importance of scientific rigorousness in evaluations.

However, it is also important to note here that methodological challenges in evaluation designs often are largely related to the complex nature of the evaluation context, which is emergent, dynamic and unpredictable (Chaplowe, Castleman & Cho, 2021). Undeniably, this ever-changing context makes a rigorous scientific approach to evaluations extremely challenging and resource intensive. Moreover, programme and intervention baseline data, with control groups, are rarely available or incomplete, making it hard to retrospectively evaluate or identify the active ingredients of change. Limited access to representative data, as in random samples of local populations/right-holders and key informants, are also often compromising data quality and the generalizability of evaluation findings. Overall, this inherent complexity in the context of our work needs to be considered if or when we see that the full learning potential of evaluations are not extracted.

We learn the most from our evaluations when they are conducted by us?

Learning is a participatory and dynamic process, intertwined with the evaluation process itself. In other words, when we use external consultants to carry out evaluations for us, the consultants might gain the most learning from the experience, and several opportunities for knowledge harvesting and internal capacity building can be lost. Considering that approximately 78% of the evaluations done in 2020 were conducted using external support (as reported in NCA's annual evaluations report), a reflection on the use of external consultants is warranted here.

Often, donors call for the use of external consultants or evaluators when conducting evaluations. This is due to the common belief that the use of external evaluators/consultants provides the objectivity and independence needed to uphold accountability and reduce bias in reporting in much the same way as an authorized accountant or accredited auditor would provide objectivity and transparency in accounting and financial audits. To some extent this belief is valid, but as discussed below, the use of external consultants might not be the panacea for bias and selective reporting donors are looking for. In fact, in some of the interviews informants reported that consultants were the cause of such bias, rather than the factor

protecting against it. Several factors seem to be at play here.

First and foremost; consultants are humans, just as prone to intentional and unintentional bias as everyone else. And many of the same forces influencing those who commissioned the evaluation, also influences the consultants. In the end both parties, consciously or unconsciously, want a report that portray pleasing results. Consultants can also feel pressured towards presenting more favourable results and make adjustment in the report that paint a more favourable outcome of the evaluation.

Another concern that arose in the interviews was the lack of high-quality consultants and the many pitfalls surrounding the procurement/contracting of consultants. In many countries, high level consultants from academic backgrounds were hard to find. There was also very little time to score bids and quality assure inception reports. Often HO support was limited when evaluating bids and there was a potential lack of local CO capacity to do so. In sum, this can often lead to the contracting of a low-quality consultant, that in turn delivers a low-quality report. Finally, some of the informants alluded to the fact that once procurement was over and the consultants had been contracted, there was little room for course correction, even when it became obvious that the consultant was not fit for the job. In sum, this hampers evaluations trustworthiness and the limits the potential for learning.

An alternative for donors might therefore be to instead focus on the credibility of the evaluations. In other words, an evaluation should not be regarded as unbiased and trustworthy just because an external consultant conducted the data collection and wrote the report – an evaluation's trustworthiness and objectivity should be judged based on its scientific quality and rigor, and the measures used to quality assure that rigor. Moreover, if the relative 'independence' of the evaluation unit within the organization is high, this will add an even higher credibility to the evaluation process. Developing systems that create a balance between the need for credibility, learning and accountability in NCA, should therefore be made a priority. While third-party generated accountability is important, it will not improve the lives of right-holders unless the information gathered is credible and used to refine our programmes. Ideally, there should be no contradistinction here, as processes securing both credibility, accountability and learning can be embedded in the evaluation process.

We learn the most from our evaluations when there is a culture for doing so?

In a previous overview of NCA's annual progress reports, we found that evaluations at NCA in general are used as accountability measures, i.e., as a way of documenting results and progress for donors. In other words, most progress reports site specific evaluation results that are valuable to donors, but these results are often not translated into specific learning outcomes that can benefit the larger organization and/or other similar programs in the future. When interviewing informants, this observation was also made as we explored findings related to learning barriers.

It seems that the main challenge in improving learning from evaluations and in turn the effectiveness of our programmes, lies not just in capturing better data through more scientific evaluations that are then widely shared, but also in the process of enabling a culture of learning in NCA. In order for evaluations to be really useful, they need to be read, studied, reviewed and discussed, and their findings must be incorporated into policies and programmes. Reviewing evidence, reflecting on its relevance and potential implications while agreeing on how to improve an intervention or programme going forward of course takes time. Making learning an integral part of our work can therefore become a bit of a challenge, and busy schedules might get in the way of extracting learning from evaluations.

In other words, high quality evaluations alone cannot create a results-based, data-driven NCA. In order to extract the most learning from evaluations, our evaluations need to be complemented by a culture for learning. This is closely related to the process of removing barriers. The insufficient use of evaluations can, as previously mentioned, be caused by high workloads that limit motivation and available resources for evaluation follow-up. In sum, removing barriers and creating a culture for learning could therefore be thought of as two important keys to unlocking NCA's full learning potential.

Recommendations

The aim of this evaluation was to identify when we learn the most from our evaluations. Findings from the present evaluation should help NCA move in the direction of a more a data-driven and results-based organization, utilizing learnings from this evaluation. In addition, findings should be used to improve the general quality of evaluations and refine NCA's systems and routines for learning reintegration.

Several key findings were identified. In response to this, the following main recommendations are suggested as priorities in relation to policy, routines and capacity building.

Improve management response letter development and follow-up

In order to improve management response letter development and follow-up, we recommend making changes in the following order:

Policy development

Revise NCA's Evaluation and Research policy and chapter 10.3 in the Operations manual. Changes should reflect the importance of the management response letter with revised routines for its use. The management response letter template is also in need of revision. The use of the new template should be made mandatory and the letter itself should include more mandatory elements and detailed instructions on how to complete it. The transformative part of the PATH-framework should also include an added emphasis on the use of management response letters in the pursuit of our evaluations' true transformative potential.

Adjustment in routines

There is a need to build new routines for management response follow-up according to the deadlines set in the management response letter. This should be done by utilizing the full potential of PIMS, building a separate evaluation workflow that correspond to mandatory elements in the evaluation process. The workflow, with prespecified tasks, should be utilized to issue reminders to identified staff, including instructions for follow-up of specific tasks and deliverables.

Capacity building

There is a need to build capacity in the completion of management response letters and how to translate evaluation findings into specific, actionable tasks that can be implemented in a prioritized order, including clear deadlines and a division of labour and responsibility. The importance of NCA's management response letter and its use should also be made known to everyone through a new launch and roll out.

Focus on context and contextualization

In order to increase contextual understanding and the contextualization of evaluations, we recommend making changes in the following order:

Policy development

Evaluations must, to a larger degree, take local context into consideration and field visits should be strongly recommended, including direct contact with key stakeholders and right-holders through user-involvement in key phases of the evaluation process. Include a paragraph on context and contextualisation in the revised Operations manual and Evaluation and Research Policy. The accountable part of the PATH-framework should also include an added emphasis on the importance of contextualization and user involvement when establishing accountability.

Adjustment in routines

A shift in routines towards more internally led evaluations should be implemented. Internally led evaluations will ensure better understanding of the local context and secure contextualization of evaluation recommendations. When external consultants are used, we need to ensure that they are familiar with the local context and that they conduct field visits when possible. NCA's Evaluation report and recommendations template is in need of revision. Focus should be on learning, and contextual considerations should be included in learning outputs to a larger degree.

Capacity building

There is a need to build capacity on the importance of contextual consideration in evaluations. Capacity building in quality assurance of bids and evaluation inception reports should also be prioritized. When giving bids a technical score, "tenderers relevant experience in the field of assignment" and "tenderers experience in the region/country e.g., knowledge of local language, culture, administrative system, government etc." should be given added weight in order to reflect the emphasis on context and contextual knowledge.

Improve timing and commitment to learning in evaluations

In order to improve the timing of evaluations, we recommend making changes in the following order:

Policy development

Revise NCA's Evaluation and Research policy and Operations Manual, with an increased emphasis on the timing of evaluations and the utility of mid-term evaluations in learning exercises. A reduction in the number of evaluations conducted each year and the cancelation of evaluations that focus on short, one-year

projects where there are no donor requirements for an evaluation, should also be discussed.

Adjustment in routines

NCA's midline evaluations and mid-term reviews should be prioritized when learning is to be extracted. All Terms of References (ToRs) for midline evaluations and reviews should include specific paragraphs where a clear intention for learning is expressed. How knowledge and learning outcomes from the evaluation will be implemented and reintegrated in the programme being evaluated and in the wider organization should also be stated.

Capacity building

Midline evaluation and mid-term reviews should to a larger degree be conducted internally in order to ensure high level of staff involvement and knowledge harvesting. It is therefore a need to build capacity in internal evaluation management and learning outcome dissemination through internal workshops and seminars. Management must also make sure there is time/room to reflect on, implement and disseminate findings from evaluations in staff schedules, incorporating e.g., reflective sessions on learning as an individual PDR goal, while actively monitoring and adjusting individual staff workload.

Reduce barriers and enable a culture for learning and knowledge reintegration

In order to decrease the number of barriers to learning from evaluations, we recommend making changes in the following order:

Policy development

NCA needs to build a stronger evaluation and learning culture and develop a shared understanding of the purpose of evaluations in meeting our learning goals. The PATH-framework's "transformative" section should be revised to reflect this and help build a culture for learning, enhancing staff understanding of the importance of learning, evidence, and the scientific approach to effective development and humanitarian aid.

Adjustment in routines

A culture for learning can also be enabled by systematically and routinely feeding lessons learnt back into organizational planning and decision-making, e.g., via regular updates from GERA in DIP-meetings. Webinars and a wide circulation of the Evaluation and Research Bulletin as an "easy to access" format would also strengthen such a culture. All NCA's evaluations should be made widely available and searchable for everyone through a dedicated Learning site containing an Evaluations library. This will make it easier to extract learning from evaluations and get an overview of the knowledge previously gathered in a specific thematic area or country.

Capacity building

There is a need to develop staff's capacity to produce credible, scientifically rigorous evaluations internally and/or in collaboration with external consultants. Capacity building in the use of a new PIMS evaluation workflow, quality assurance of inception reports, reports and recommendations should also be prioritized. This will reduce barriers to learning and enable staff to extract learning from evaluations and the evaluation process itself more fully. Staff involved in evaluation and/or programme development should routinely extract learning from previously conducted evaluations and documented learning processes.

Concluding remarks

In the present global evaluation, a broader understanding of when we learn the most from our evaluations has been obtained. We have also identified several important areas where there is a large potential for improvement in order to enhance learning. With this understanding, we can now begin to build a best practice framework for evaluation processes and knowledge reintegration in the organization through adjustment in policy, routines and capacity building.

Yet, in order for this report to be fully utilized, we need to create a sense of urgency for learning and the implementation of the recommendations presented. Following our own recommendations, this evaluation should therefore be made widely available, ensuring widespread uptake and utilization of findings. A specific and feasible management response letter with clear deadlines and division of labour/responsibility should also be produced, including plans for later follow-up.

There is also a need to expand upon the data and results from the present evaluation in a longitudinal design, with a follow-up in 2023. By doing so, we will be able to identify whether the recommendations prescribed in this report have been implemented, and whether the implementation of those recommendations have produced the desired results.

References

Cracknell, B. E. (2000). Evaluating development aid: issues, problems and solutions. Sage.

Chaplowe, S., Castleman, AM. and Cho, M. (2021). Evolving evaluation practice: past, present and future challenges. ALNAP.

Appendix I

Method and results

Data sources and scope

Data sources and scope

The geographical scope of the evaluation was global and included a retrospective desk review of evaluations, cross-sectional survey, and key informant interviews.

Quantitative method

The sample frame consisted of all evaluations conducted in the last 2 years (2018 and 2019) and their related management response letters. In total 25 evaluations and their respective recommendations and management response letters were identified.

Measures

Survey: A survey consisting of both quantitative and qualitative measures were distributed to HO and CO staff who were involved in the selected evaluations. Survey questions included items concerning overall quality of methods, reports, recommendations and management response letters (5 items), items concerning value of recommendations and management response letters (2 items) and items concerning degree of implementation, learning and reintegration of knowledge (3 items). Finally, three items concerning overall satisfaction with the evaluation and its learning outcome was also included.

Quality Ratings: All included evaluations were rated in relation to the quality of the recommendations presented and the corresponding management response letter by three independent raters in MEL. All separate recommendations were rated on a scale from 1-5 in relation to specificity and feasibility, while management response letters were rated in relation to whether they included specific, actionable responses, a clear division of labour and specified timelines/ deadlines for implementation. The ratings from each rater were summarized and averaged for each evaluation and evaluations in general.

Statistical analyses

All preregistered questions were analyzed using IBM SPSS statistical software package, version 28. Descriptive statistics based on survey responses were performed to summarize the "Overall evaluation quality and value for future programming", the "Value of evaluation recommendations and management response action points", "Implementation, learning and reintegration of knowledge", and finally, the "Overall satisfaction with the evaluation and its outcome". Associations between quality predictors and learning outcome dimensions were investigated using bivariate correlation matrixes with Pearson Correlation Coefficients.

Qualitative method

Ten evaluations were selected for in-depth, semi-structured interviews with key informants based on HO ratings (5 high and 5 low) of the evaluations' recommendations and management response letters. Key informants include CO staff directly involved in planning of evaluations and/or the development of management responses.

The interview guide consisted of 16 questions corresponding to four main themes. The themes covered were "Overall quality of the evaluation", "the Evaluation Process", "the Evaluation Response", and general questions regarding "the Usefulness of the evaluations and NCA's approach to evaluations".

The interviewers, NCA's Global Evaluation and Research Advisor (GERA) and NCA's Senior advisor for Learning, first discussed the content of the interviews and the interview process. During these discussions 5 main themes and 43 sub-themes for investigation were identified. The transcripts of the Key informant interviews (KII) and qualitative data from the survey questions (items 9, 11, 13 and 16) were then imported into and analysed in NVivo Qualitative Analysis software, version 12. In NVivo, Interviews and survey data were coded according to the 48 metrics that were manually created. Analyses were done to cross reference coding nodes to identify trends and test hypotheses from the preregistration stage. Key citations from interviews were identified using NVivo, while survey items that illustrate findings were identified manually by GERA.

At the analysis stage, data on "usefulness" was cross-referenced with evaluation quality themes to test whether there was a positive association between subjective quality and perceived usefulness of evaluations. The three items from the online survey focused more on why recommendations/action points were not implemented/utilized or produced learning, and why the knowledge/learning gained were not reintegrated going forward, and was therefore presented separately in the report.

Quantitative results

In this section we describe the details of the quantitative results of the evaluation. In total, we received 26 survey responses related to 20 evaluations. 5 evaluations did not receive any responses and were therefore excluded from the results.

CO staff's ratings of the 20 evaluations are presented in table 1-4.

Table 1. *Rating of overall evaluation quality and value for future programming.*
(1=Very poor, 2=Poor, 3=Acceptable, 4=Good, 5= Very good)

Question	Mean	Min	Max	Std. Deviation
How would you rate the overall quality of the methodology?	4.09	3	5	.66
How would you rate the overall quality of the final report?	3.91	3	5	.62
How would you rate the overall quality of the summary and recommendations?	4.03	3	5	.68
How would you rate the overall quality of the management response letter?	3.93	2	5	.83

Table 2. *Value of evaluation recommendations and management response action points.*
(1=No value, 2=low value, 3=Moderate value, 4 High value, 5= Very high value)

Question	Mean	Min	Max	Std. Deviation
Where the recommendations presented in the summary of value for future programming?	3.93	3	5	.77
Where the action points presented in the management letter of value for future programming?	3.93	2	5	.87

Table 3. *Implementation, learning and reintegration of knowledge.*
(1=Not at all, 2=To a low degree, 3=Moderate degree, 4=High degree, 5= A very high degree)

Question	Mean	Min	Max	Std. Deviation
To what degree do you feel that the evaluation produced actionable and relevant learning outcomes that could be integrated in the future?	3.83	2.5	5	.77
Overall, how satisfied are you with the evaluation process and its outcomes?	4.04	3	5	.68

Table 4. *Overall degree of learning and satisfaction.*
1= Not at all, 2=To a low degree, 3=Moderate degree, 4=High degree, 5=To a very high degree
1=Very dissatisfied, 2=Dissatisfied, 3=Neither satisfied nor dissatisfied, 4= Satisfied, 5= very satisfied

Question	Mean	Min	Max	Std. Deviation
To what degree was the recommendations/ action points in the management response letter implemented in future programming?	3.21	1	5	.92
To what degree did the recommendations/action points in the management response letter produce learning for future programming?	3.55	2	5	.95
To what degree were the knowledge gained from the evaluation reintegrated in other areas of the organization?	3.20	2	5	.77

Overall quality rating from the MEL-team

As described in this report, all included evaluations were rated in relation to the quality of the recommendations presented (specificity and feasibility) and the corresponding management response letter (specificity, division of labor and deadlines) by three independent raters in MEL. The ratings from each rater were summarized and averaged for each evaluation and evaluations in general.

Results revealed that raters from the MEL-team overall rated the quality of recommendations and the management response letters slightly lower than CO staff, with MEL means of 3.56 and 3.32 respectively versus 4.03 and 3.93 for CO staff (See table 1-4 for CO staff ratings). The variance in MEL-ratings was also larger with recommendations and management responses receiving ratings of 1-5. Finally, MEL-raters reported that most recommendations and management responses were hard to rate due to their brevity and general lack of detail and use of generic language, which could be considered a finding in itself.

Associations/correlations

Our aim for these of analyses, was to explore the potential strong association between the quality of evaluation methodology (design, data collection and analysis), the quality of the evaluation report (clear, easy to read, of appropriate length and useful), and learning/knowledge reintegration in NCA. As described in the preregistration form, our main hypothesis was that high levels of learning outcome utilization/ implementation, learning for future programming, and finally, knowledge/learning reintegration in other areas of the organization would be predicted by specific, actionable and feasible recommendations and management response letters, rooted in high quality evaluations.

Our hypotheses and a priori line of thinking in relation to this, is summarized in figure 1.

When analyzing associations between these factors, a somewhat more complex picture emerged, as shown in figure 2 below.

When analyzing associations between quality predictors and learning outcome dimensions using bivariate correlation matrixes, the variable "overall quality of the management response" was positively associated with "the degree to which recommendations/action points presented in the management response letter produced learning for future programming" and "the degree to which the knowledge/learning gained from the evaluation was reintegrated in other areas of the organization", with moderate correlations ($r(18) = .56$, $p < .01$ and $r(18) = .51$, $p < .02$, respectively), as shown in figure 2.

However, the "overall quality of the evaluation summary and recommendations" were not associated with "the degree to which recommendations/action points presented in the management response letter produced learning for future programming" ($p=.74$) or "the degree to which the knowledge/ learning gained from the evaluation was reintegrated in other areas of the organization" ($p=.56$), marked as .ns in figure 2. Finally, the quality of recommendations and management responses were not significantly associated with the "degree to which recommendations/action points in the management response letter were implemented/utilized in future programming" ($p=.47$ and $p=.28$, respectively).

Figure 1.

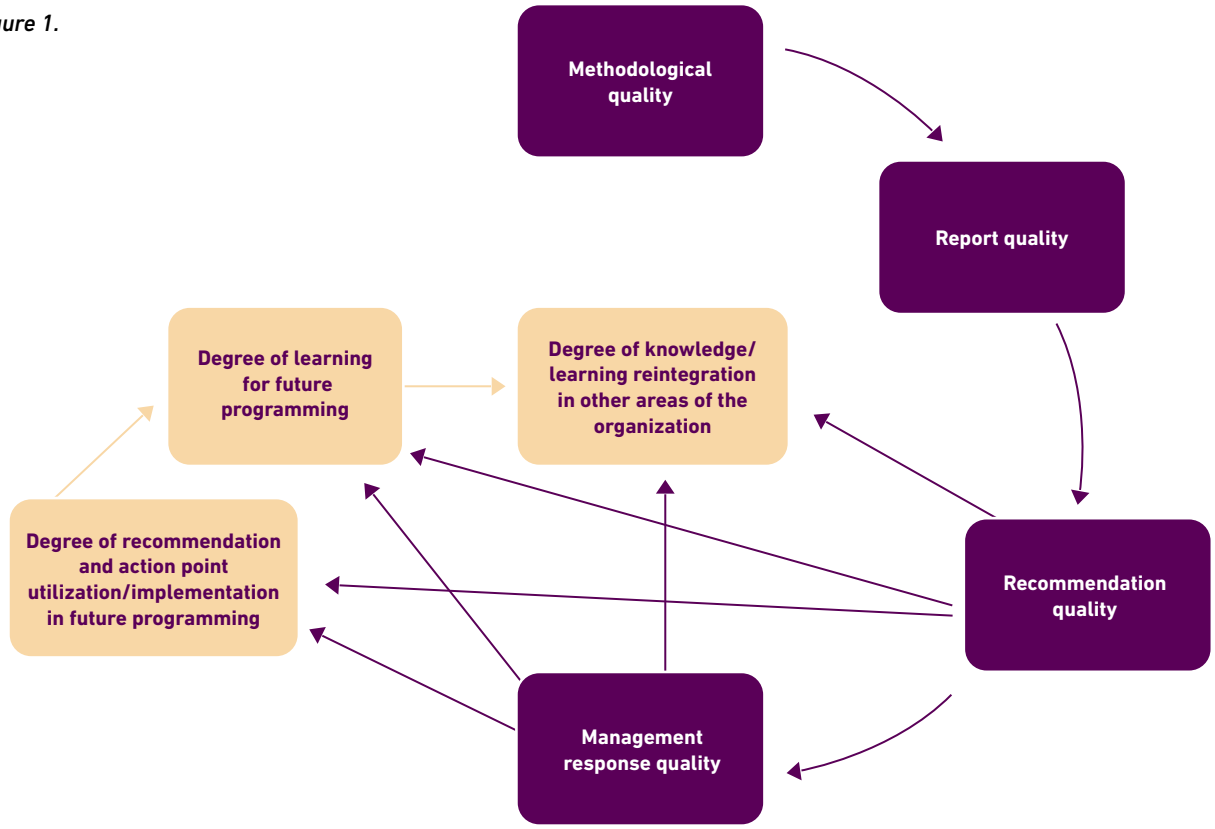
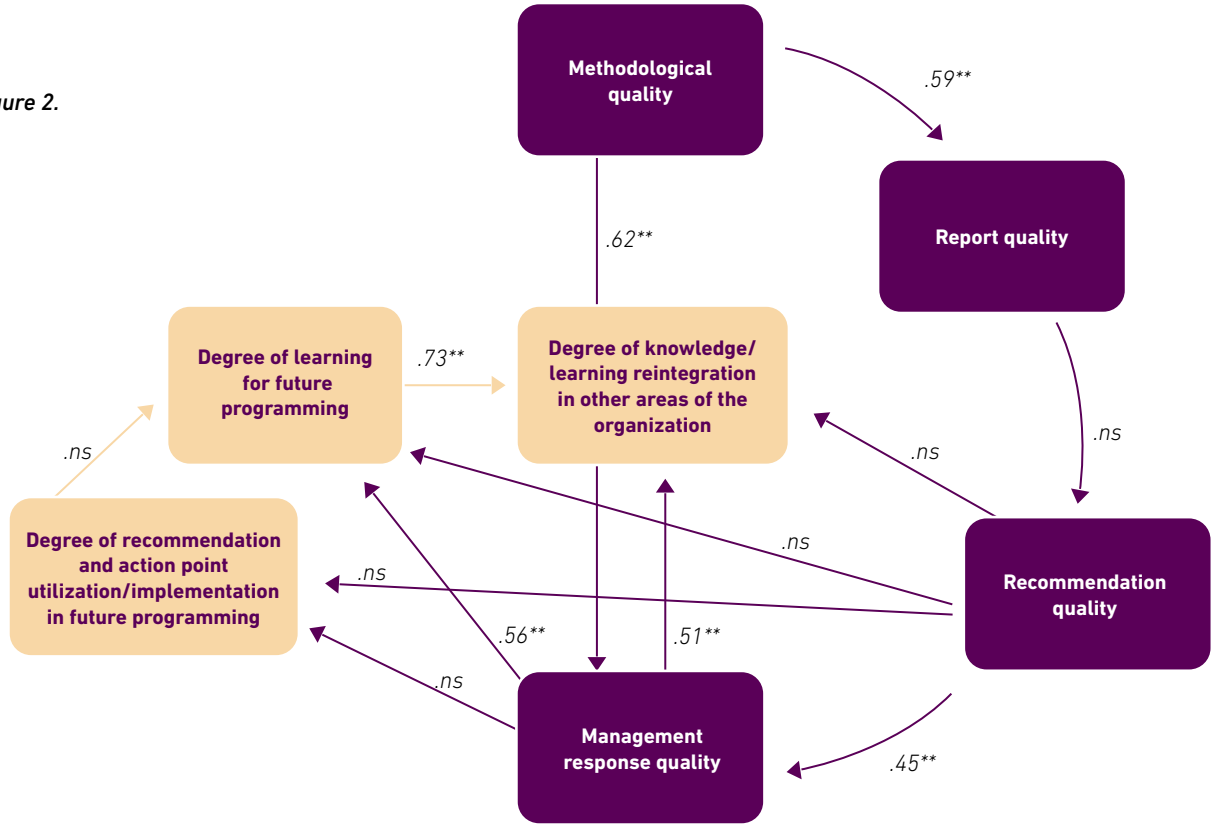


Figure 2.



Appendix II

Survey questionnaire

When do we learn the most from our evaluations? A survey among NCA staff

Due to your involvement in an evaluation conducted in 2018 or 2019, you have been selected as a valued respondent for this survey! The survey focuses on what we learn from evaluations. Learning is here defined as:

"...the process of acquiring a new understanding and/or gaining new knowledge as a result of evaluations that, in turn, improves future programming."

NB. The survey relates to a specific evaluation that you participated in, as specified in section 1 below. Please answer all questions as honest as possible.

Section 1

Quality of evaluation

1.Please select the evaluation that you are rating from the list below. Keep this evaluation in mind as you answer the remaining questions in this survey.
Select your answer

2.How would you rate the overall quality of the evaluation methodology (design, data collection and analysis)?
1=Very Poor, 2=Poor, 3=Acceptable, 4=Good, 5=Very Good

3.How would you rate the overall quality of the final report (clear, easy to read, of appropriate length, and useful)?
1=Very Poor, 2=Poor, 3=Acceptable, 4=Good, 5=Very Good

4.How would you rate the overall quality of the summary and recommendations (specific, actionable and feasible)?
1=Very Poor, 2=Poor, 3=Acceptable, 4=Good, 5=Very Good

5.How would you rate the overall quality of the management response letter (specific, actionable and feasible)?
1=Very Poor, 2=Poor, 3=Acceptable, 4=Good, 5=Very Good

Section 2

Value of evaluation recommendations and management response action points

Please rate the following items with the evaluation summary and management response letter in mind

6.Were the recommendations presented in the evaluation summary of value for future programming?
1=No value, 2=Low value, 3=Moderate value, 4=High value, 5=Very high value

7.Were the action points presented in the management response letter of value for future programming?
1=No value, 2=Low value, 3=Moderate value, 4=High value, 5=Very high value

Section 3

Implementation, learning and reintegration of knowledge

Please rate the following items with the specific evaluation outcomes in mind

8.To what degree were the recommendations/action points presented in the management response letter utilized/implemented in future programming?
1= I do not know, 2=Not at all, 3=Low degree, 4=Moderate degree, 5=High degree, 6=To a very high degree

9.If recommendations/action points were not implemented/utilized to a high degree (a rating of 5 or 6), why was this so?
Please fill in your answer.

10.To what degree did recommendations/action points presented in the management response letter produce learning for future programming?
1= I do not know, 2=Not at all, 3=Low degree, 4=Moderate degree, 5=High degree, 6=To a very high degree

11.If recommendations/action points did not produce learning to a high degree (a rating of 5 or 6), why was this so?
Please fill in your answer.

12.To what degree were the knowledge/learning gained from the evaluation reintegrated in other areas of the organization?
1= I do not know, 2=Not at all, 3=Low degree, 4=Moderate degree, 5=High degree, 6=To a very high degree

13.If the knowledge/learning gained were not reintegrated to a high degree (a rating of 5 or 6), why was this so?
Please fill in your answer.

Section 4

Overall satisfaction with the evaluation and its outcome

14.To what degree do you feel that the evaluation produced actionable and relevant learning outcomes that could be reintegrated in the future?
1=No degree, 5=To a very high degree

15.Overall how satisfied are you with the evaluation process and its outcomes?
1=Very dissatisfied, 2=dissatisfied, 3=neither satisfied nor dissatisfied, 4=satisfied, 5 stars=Very satisfied

16.In your opinion, what would make NCA's evaluations better and more useful in the future?
Please fill in your answer.

Appendix III

Interview guide

"Thank you for making the time for this interview. Because of your involvement with X Evaluation, you have been identified as a key informant for our wider study of When We Learn Most from Our Evaluations. The purpose of this study is to improve the quality and usefulness of evaluations across NCA. We are interested to hear your honest impressions and feedback on X Evaluation, as well as your inputs on evaluations in NCA more generally. Your responses will be anonymous."

Could you give us some more information as to what makes this evaluation particularly good/bad for you? For example, was the methodology good/bad; Was it relevant to your work?; Overall quality of the final report?; Were the recommendations specific and actionable?

What about the management response letter? Did it respond to the recommendations? Did it give specific timelines for the Reponses? In your opinion did the quality of the management repose affect whether or not the recommendations were addressed?

Did this evaluation feel relevant and needed? Was it initiated locally or requested by a donor?

What were the main learning outcomes from the evaluation and how did these outcomes affect your work going forward? Did you change anything due to the findings? Or just a good learning experience to have in mind going forward?

Could you describe the typical process at your office for beginning a new evaluation? I.e., assessing proposals, data collection, etc.

When you think about the process for this evaluation, was it typical of how your office normally operates? If yes/no, how did that impact the quality of the evaluation if at all?

Do you feel that you personally and your office generally have the tools and skills you need to evaluate proposals from consultants or develop the ToR and Inception report?

When you receive a number of proposals, do you feel comfortable assessing which is the highest quality? What sort of method do you use for this?

Did you feel that you personally and your office had the tools and skills needed to quality assure the evaluation report for this evaluation? When you received the draft evaluation report, did you feel comfortable assessing its quality and offering ways to improve it?

What is the typical course of action at your office once an evaluation has been completed? What procedures are in place following an evaluation? For example, management response, team retrospective meetings, etc.

When you think of the evaluation response for this evaluation was it typical of how your office normally operates? If yes/no, how did this impact the degree to which the evaluation recommendations were addressed?

Were the knowledge/learning gained from this evaluation as a whole reintegrated in the wider organization? If yes, how was this done? If not, why? Any obstacles?

In your opinion, when are evaluations THE MOST useful? When evaluations are useful, what factors make them useful to you?

In your opinion, when are evaluations THE LEAST useful? When evaluations are not useful, what factors hinder usefulness?

In your opinion how should evaluations be used? What should we change about our approach to evaluations to make sure that we get the most out of them?

Is there anything else you would like to add about either this evaluation in particular or evaluations more generally?

Appendix IV

Preregistration

Preregistered: 11.06.21, 13:21 CET

NORWEGIAN CHURCH AID
actalliance

Preregistered

Preregistration Form

Name of intervention to be evaluated: NCA's evaluations.

Name of Evaluation /research project: When do we learn the most from evaluations?

Baseline, midline or endline, other? Baseline measure

Short evaluation description:
The purpose of this evaluation is to establish a baseline of evaluation quality and knowledge reintegration in NCA, while at the same time exploring potential obstacles and/or synergies that might hamper or increase learning. The overall aim is to improve evaluation quality, management response quality, uptake and reintegration of lessons learned in the organization.

Our focus will be to explore how and to what extent the quality of evaluations and management responses predict learning, and in turn change and/or improve NCA's work. The aim here is to identify best practice cases, while at the same time investigate potential obstacles, pitfalls and common mistakes that prevent learning from taking place, and finally how to correct for them in the future.

DAC-criteria investigated:
- Coherence
- Effectiveness
- Impact

1. **Have any data been collected for this evaluation already?**
   - Yes, some parts of the data set have been collected already.
   - Some data still needs to be collected.
   - No analyses have been conducted.

2. **What is/are the main question(s) being asked?**

   Main questions:
   - What is the general quality of evaluations, evaluation recommendations and management response letters?
   - Does the quality of recommendations and management responses predict degree of implementation, learning and knowledge reintegration?

   Key sub questions:
   - When evaluations are used in a constructive/transformative way, what brings this about?
   - When evaluations are not used in a constructive/transformative way, what is hampering the use?
   - How can we enhance the degree of implementation, learning and knowledge reintegration?

3. What are the main hypotheses being tested (be specific)?

Main hypothesis:
After an evaluation is conducted, high levels of learning outcome implementation, learning and knowledge reintegration in the organization is predicted by specific, actionable and feasible recommendations and management response letters that include specific, high quality action points, a specified timeline, and a clear division of labor/responsibility.

Hypothesis breakdown:

A. NCA evaluations that are a priori planned and designed with the specific intent of learning, produce the most actionable recommendations and learning outcomes. In turn, this produces high-quality management responses that are easily implemented in future programming and/or policy development.
B. If, on the other hand, evaluation recommendations are vague/generic/non-specific/not actionable/not feasible, clear learning outcomes are limited.
C. More specifically, a lack of clear learning outcomes and specific, actionable recommendations hamper the development of management response quality. When presented, management responses are therefore also non-specific, vague, and generic in their form, with non-explicit timelines, no clear division of labor/responsibility and few concrete action points.
D. This lack of specific deadlines, division of labor/responsibility, specific recommendations and concrete action points, create high levels of responsibility diffusion and unclear goals/targets.
E. Finally, this makes implementation of action points difficult, and implementation, learning and reintegration of knowledge is diminished or postponed indefinitely.

4. Describe the key indicator(s) specifying how they will be measured:

Observational data (n=25 evaluations)

Background variables/index variables:

- Evaluation number: 1-25
- Year of completion: 2018-2019
- Number of recommendations: 1-15
- MRL response: Y/N
- Global programme: GBV, CRWASH, PEACE, Other?
- Initiator: HO or CO?
- External, internal or mixed evaluation team?
- Baseline, Midline/Endline/Other?

Quality measures:

- Specificity/actionability in recommendations (Mean and standard deviations, range 1-5).
  - 1=No specificity or actionable recommendations
  - 5=All recommendations are actionable with a high level of specificity

- Feasibility of recommendations (Mean and standard deviations, range 1-5).
  - 1=Unfeasible recommendations
  - 5=Highly feasible recommendations

- Specificity/actionability in Management Response Letter (Mean and standard deviations, range 1-5).
  - 1=No specificity or actionable responses
  - 5=All responses are actionable with a high level of specificity

- Division of labor/responsibility in Management Response Letter (Mean and standard deviations, range 1-5).
    - 1=No mention of division of labor/responsibility in responses
    - 5=All responses are linked to specified names/functions with a clear division of responsibility/labor

- Deadlines/timelines in Management Response Letter (Mean and standard deviations, range 1-5)
    - 1=No mention of timeline or deadline for responses
    - 5=All responses have clear deadlines

Survey data (n=25-30 respondents)

Quality measures:
How would you rate the quality of the evaluation (design, data collection and analysis)?
- 5=Very Good, 4=Good, 3=Acceptable, 2=Poor, 1=Very Poor

How would you rate the overall quality of the final report?
- 5=Very Good, 4=Good, 3=Acceptable, 2=Poor, 1=Very Poor

How would you rate the quality of the summary and recommendations?
- 5=Very Good, 4=Good, 3=Acceptable, 2=Poor, 1=Very Poor

How would you rate the quality of the management response letter?
- 5=Very Good, 4=Good, 3=Acceptable, 2=Poor, 1=Very Poor

Measures of value and relevance of recommendations/action points:
Where the action points presented in the management response letter of value for the future programming?
- 5=Very high value, 4=High value, 3=Moderate value, 2=Low value, 1=No value

Where the recommendations presented in the management response letter relevant for the future programming?
- 5=Very high relevance, 4=High relevance, 3=Moderate relevance, 2=Low relevance, 1=No relevance

Measures of implementation, learning and reintegration:
To what degree were the recommendations/action points presented in the management response letter utilized/implemented in future programming/interventions?
- 5=To a very high degree, 4=High degree, 3=Moderate degree, 2=Low degree, 1=Not at all
- If not to a high degree, why?

To what degree did recommendations/action points presented in the management response letter produce learning for future programming/interventions and policy development?
- 5=To a very high degree, 4=High degree, 3=Moderate degree, 2=Low degree, 1=Not at all
- If not to a high degree, why?

To what degree were the knowledge/learning gained from the evaluation as a whole reintegrated in the wider organization?
- 5=To a very high degree, 4=High degree, 3=Moderate degree, 2=Low degree, 1=Not at all
- If not to a high degree, why?

Key informant data (n=5-10 respondents):

Open-ended questions concerning quality, implementation, learning and reintegration. Case stories.

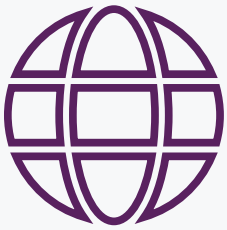
5. **Specify evaluation design and type data collection (e.g., quasi-experiment, survey, focus group, desk review):**
   - Design: Cross-sectional, retrospective.
   - Data collection: A combination of desk-review/observational data, survey and key informant interviews.

6. **Specify which key analyses you will conduct to examine the main question/hypothesis:**

   - Descriptive analyses of all measures (Means and Standard deviations) presented in tables and graphs.
   - Correlation matrix for all measures with Pearson correlation coefficient
   - Analysis of Variance (ANOVA) with Bonferroni post-hoc tests;
     - Independent variables (grouping variable): low, mid and high-quality Evaluations
     - Independent variables (grouping variable): low, mid and high-quality Management responses
     - Dependent variables: Degree of implementation
     - Dependent variables: Degree of learning
     - Dependent variables: Degree of organizational reintegration

7. **How many observations will be collected and what will determine the sample size?**

Sample frame is determined by the number of evaluations conducted in 2018 and 2019. A random sample of 50% of these evaluations will be evaluated/rated. Number of survey informants and key interviews are determined by finding in the evaluations, but are estimated to be n=25-30 and n=5-10, respectively.



Global Evaluations
and Research

Norwegian Church Aid's Global Evaluations and Research programme

Norwegian Church Aid's Global Evaluations and Research programme helps ensure that our humanitarian and development assistance is as relevant, coherent, efficient, effective, sustainable, and impactful as possible.

The main focus of Norwegian Church Aid's Global Evaluations and Research Programme is to gather evidence and identify higher-level learning outcomes that can be broadly implemented in our future programmatic work.

www.nca.no
E-mail: nca-oslo@nca.no
Telephone: +47 22 09 27 00 Fax: +47 22 09 27 20
Street address: Bernhard Getz' gate 3, 0130 Oslo, Norway

Account no: 1594 22 87248

NORWEGIAN CHURCH AID
actalliance