META-EVALUATION ON THE QUALITY OF (PROJECT) EVALUATIONS IN GERMAN DEVELOPMENT COOPERATION

2022

DEval
GERMAN INSTITUTE FOR DEVELOPMENT EVALUATION

The cross-organisational meta-evaluation examines project evaluations of eleven governmental and non-governmental organisations in Germany that were (co-)funded by the BMZ. It analyses their understanding of quality in evaluations and their application of internationally recognised quality standards, especially the OECD-DAC and DeGEval standards. It also analyses factors linked to the application of the quality standards. Overall, the meta-evaluation shows that the application of the quality standards examined is largely part of the established evaluation practice of the organisations involved. Beyond that, however, these are not yet consistently or systematically prescribed or clearly documented in the organisational documents and processes or at the level of the individual evaluation. It was not possible to identify factors linked to the application of several quality standards. The meta-evaluation makes recommendations to the involved organisations. These are designed to (i) ensure that the quality standards are systematically institutionalised, and (ii) guarantee joint learning in the future through a systematic exchange of experience. It also recommends that the BMZ develop an analysis grid for application of the quality standards based on the BMZ Evaluation Policy now in force, and make this available to the organisations.

META-EVALUATION ON THE QUALITY OF (PROJECT) EVALUATIONS IN GERMAN DEVELOPMENT COOPERATION

2022

IMPRINT

Authors

Dr Kerstin Guffler
Laura Kunert
Marian Wittenberg
Dr Nico Herforth

Responsible team leader

Dr Kerstin Guffler

Responsible head of department

Amélie Gräfin zu Eulenburg

Cover and graphic design

Katharina Mayer

Editing

Marcus Klein, PhD

Translation

Dr John Cochrane

Photo credits

Cover: VektorMine, Shutterstock

Bibliographical reference

Guffler, K., L. Kunert, M. Wittenberg and N. Herforth (2022), *Meta-evaluation on the Quality of (Project)Evaluations in German Development Cooperation*, German Institute for Development Evaluation (DEval), Bonn.

ISBN 978-3-96126-191-8 (PDF)

Published by

German Institute for
Development Evaluation (DEval)
Fritz-Schäffer-Straße 26
53113 Bonn

Phone: +49 (0)228 33 69 07-0
E-mail: *info@DEval.org*
www.DEval.org

The German Institute for Development Evaluation (DEval) is mandated by the German Federal Ministry for Economic Cooperation and Development (BMZ) to independently analyse and assess German development interventions.

The Institute's evaluation reports contribute to the transparency of development results and provide policymakers with evidence and lessons learned, based on which they can shape and improve their development policies.

This report can be downloaded as a PDF file from the DEval website:
https://www.deval.org/en/publications

A BMZ response to this evaluation is available at *https://www.bmz.de/de/ministerium/evaluierung/ bmz-responses-19422*

ACKNOWLEDGEMENTS

When working on the meta-evaluation of the quality of (project) evaluations in German development cooperation, the evaluation team received support from various stakeholders. At this point we would like to thank them for their support.

We express our gratitude to the *members of the reference group*. This comprised representatives of the BMZ, and the evaluation units/desks of the involved organisations – specifically the Federal Institute for Geosciences and Natural Resources, CARE Germany, the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, the German Red Cross, the Institute for International Cooperation of the German Adult Education Association, the Protestant Agency for Diakonie and Development, the Heinrich Böll Foundation, KfW Development Bank, the Konrad Adenauer Foundation, MISEREOR and the National Metrology Institute of Germany – as well as representatives of VENRO – the umbrella organisation of development and humanitarian non-governmental organisations (NGOs) in Germany. We would like to thank them for acting as a sounding board, and for supporting us in correctly presenting the object of the evaluation, providing us with data and documents, and commenting on our work in writing and verbally. Furthermore, we would also like to thank them for the positive, constructive and appreciative dialogue we were able to hold over time.

We also thank Prof. Dr Wolfang Beywl and Prof. Dr Thomas Widmer. They accompanied and supported us throughout the entire evaluation process in their respective capacities as *external peer reviewer* and *evaluation quality expert*. This included providing written comments as well as constructive feedback in workshops on various interim products. Internally at DEval we would like to thank Dr Martin Noltze, who had already conducted the previous cross-organisational meta-evaluation. He not only made an important contribution to quality assurance as an *internal peer reviewer*, but also shared his valuable expertise and previous experience with us.

We would also like to thank selected experts for their specific support on meta-evaluations, OECD-DAC and DeGEval standards, sampling and the use of the MAXQDA software application. They provided us with important information and insights for the design of this meta-evaluation.

Last but not least, we would like to say thank you to our colleagues Dr Thomas Wencker, Jens Eger and the whole community of practice team at DEval's Competence Centre for Evaluation Methodology.

Thank you!

EXECUTIVE SUMMARY

Introduction

Meta-evaluations can be described as evaluations of evaluations. They are becoming increasingly important in development cooperation. According to Caracelli and Cooksy (2009: 2), meta-evaluations are conducted "to improve an evaluation in process, reflect systematically on the strengths and weaknesses of an evaluation to enhance one's future evaluation practice, or provide information about the credibility of the findings to users." This understanding is also adopted in the present meta-evaluation. There are now a growing number of meta-evaluations in international development cooperation. In German development cooperation too, intra-organisational and cross-organisational meta-evaluations are now being carried out.

This meta-evaluation ran concurrently with the preparation of the Evaluation Policy for German Development Cooperation, published by the Federal Ministry for Economic Cooperation and Development (BMZ). It follows on among others from previous assessments of the German development cooperation system, the study on monitoring of a system assessment, and the sustainability meta-evaluation (Noltze et al., 2018). In 1999, the BMZ commissioned a first systematic assessment of evaluation practice in German development cooperation (Borrmann et al., 1999). Ten years later, a second system assessment was completed (Borrmann and Stockmann, 2009). Building on the latter, in 2015 DEval conducted a study to monitor implementation of its findings and recommendations (Lücking et al., 2015). In 2018, DEval published a cross-organisational meta-evaluation[1] on the quality on the quality of project evaluations of the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH and KfW Development Bank (Noltze et al., 2018).

The BMZ published its Evaluation Policy for German Development Cooperation (BMZ, 2021a) prior to the completion of this meta-evaluation, thus setting the most recent milestone. This policy further consolidates the understanding of evaluation quality in German development cooperation. In particular, the policy states that both the standards of the Development Assistance Committee (DAC) of the Organisation for Economic Co-operation and Development (OECD), and the standards of the Evaluation Society in Germany (DeGEval), are binding for the official implementing organisations. It also states that they provide guidance for the non-governmental organisations. Since the period of the meta-evaluation preceded the adoption of the policy, it was not possible to include it in the analysis. Nevertheless, important findings were generated for the future unfolding of the policy.

One particular feature of this meta-evaluation is that it examines the application of quality standards in evaluations for a variety of both official implementing (governmental) and non-governmental organisations. The following organisations were involved in the meta-evaluation: the Federal Institute for Geosciences and Natural Resources (BGR), CARE Germany, the Institute for International Cooperation of the German Adult Education Association (DVV), the German Red Cross (DRK), the Protestant Agency for Diakonie and Development (EWDE), GIZ, the Heinrich Böll Foundation (hbs), the Konrad Adenauer Foundation (KAS), KfW, MISEREOR and the National Metrology Institute (PTB). This meta-evaluation examined project evaluations conducted between October 2016 and December 2022 for which German organisations were (co-)responsible. Germany's official implementing (governmental) organisations were all included. The non-governmental organisations were selected using criteria that reflected their structural heterogeneity. This enabled the findings to reflect the broadest possible range of experience with the application of quality standards.

[1] The BMZ's Evaluation Policy for German Development Cooperation states that DEval meta-evaluations are part of the quality assurance framework for the evaluation system (BMZ, 2021a).

The present meta-evaluation looked at the application of the quality criteria in relation to the requirements for the involved organisations to apply the OECD-DAC and/or the DeGEval standards. It also examined the evaluations of GIZ and KfW in relation to the quality criteria of the sustainability meta-evaluation.

This cross-organisational meta-evaluation aims to generate findings on the involved organisations' understanding of evaluation quality, and on strengths and weaknesses in the application of the quality standards. It also identifies and analyses factors linked to the application of the quality standards. To support future learning, the meta-evaluation also aims to find explanations for the non-application of quality standards. To achieve the above, the meta-evaluation addressed the following evaluation questions.

Evaluation questions

1. What understanding of evaluation quality do the involved German development cooperation organisations have?
2. To what extent are quality standards applied in evaluations of the involved German development cooperation organisations?

a) To what extent are strengths and weaknesses evident in the application of the OECD-DAC and the DeGEval standards in the evaluations of the involved German development cooperation organisations?
b) To what extent are strengths and weaknesses evident in the application of the organisation-specific quality standards in the evaluations of the involved German development cooperation organisations?
c) To what extent are strengths and weaknesses evident in the application of the sustainability meta-evaluation quality criteria in the evaluations of the GIZ and KfW?

3. To what extent are country-specific, evaluation-specific and organisation-specific factors linked to the application of quality standards?

Theoretical and empirical background

The understanding of quality

This meta-evaluation equates evaluation quality with the application of the relevant quality standards, i.e. the quality standards that are required for the involved organisations. It then examines evaluation quality accordingly. This definition of quality includes – but is not limited to – the content of the OECD-DAC and DeGEval standards documents. The basis for identifying evidence of good evaluations is provided by the OECD-DAC and DeGEval standards. This is due to their international recognition, their link to development cooperation and their relevance to German development cooperation organisations. The term "application of quality standards" was chosen in consensus with the reference group[2]. It describes the extent to which evidence can be obtained as to whether and how the quality standards were included in the evaluations examined – i.e. whether this was documented in writing, or was reported back in writing upon request by the evaluation team. The OECD-DAC and DeGEval standards are maximum standards. This means that the involved organisations do not have to apply all quality standards in all evaluations. These internationally recognised quality standards were complemented with organisation-specific quality standards and the quality criteria of the sustainability meta-evaluation.

2 The reference group comprised representatives of the involved organisations and VENRO, and officers from BMZ Division GS 22 "Evaluation and development research, DEval, IDOS". Its members accompanied the evaluation process during all phases of the evaluation (for example through virtual meetings or comments on evaluation documents DEval, 2021a).

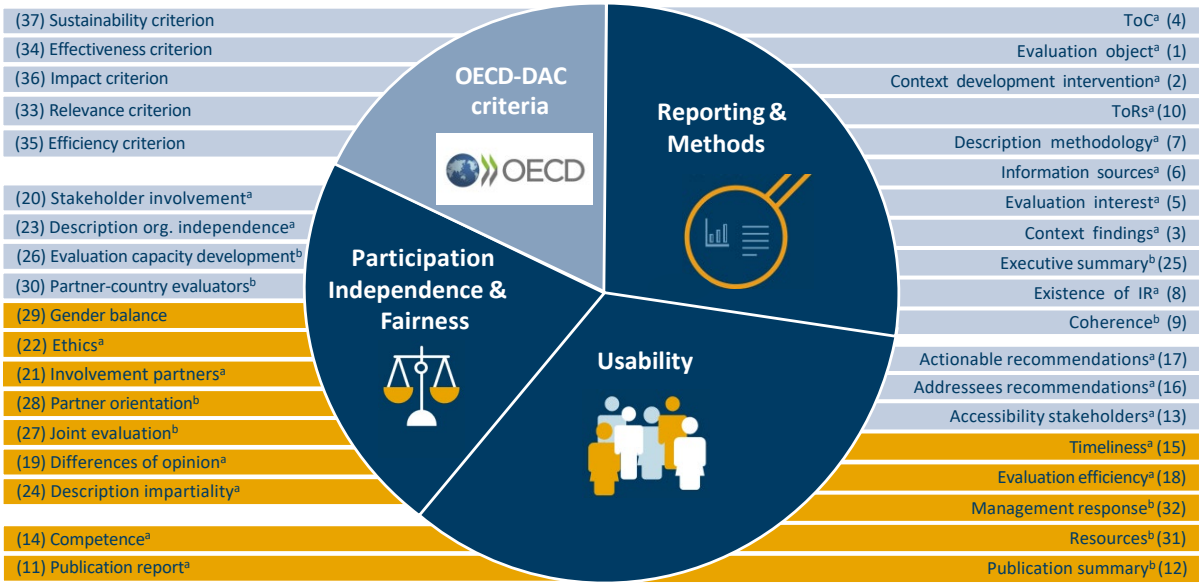
Analysis grid

The analysis grid includes quality criteria derived from the OECD-DAC and DeGEval standards documents. It also includes organisation-specific quality criteria and quality criteria from the sustainability meta-evaluation (Noltze et al., 2018). The OECD-DAC and DeGEval quality criteria can be classified into three areas. These are: 1) the overlap between the standards documents of OECD-DAC and DeGEval, 2) the OECD-DAC standards document minus the overlap with the DeGEval standards document (OECD-DAC only) and 3) the OECD-DAC criteria. Since all involved organisations were committed to the "OECD-DAC criteria" (BMZ, 2006), these are listed as a separate area, although their application is part of the OECD-DAC standards (quality standard 2.8 OECD-DAC, 2010). For DRK, EWDE, GIZ and hbs, the analysis grid also included organisation-specific quality criteria.[3] For GIZ and KfW it included quality criteria already addressed in the previous sustainability meta-evaluation. criteria already addressed in the previous sustainability meta-evaluation.

Standard clusters

With the exception of the five OECD-DAC criteria, the quality criteria were assigned to three standard clusters – "reporting and methods", "participation, independence and fairness" and "usability".[4] The standard cluster "reporting and methods" mainly comprises quality criteria that relate to either the presentation of information on the evaluation methodology, or the existence or content of selected evaluation documents (Figure 1). The standard cluster "participation, independence and fairness" is primarily assigned quality criteria that concern the inclusion of different groups of people in the evaluation. The standard cluster "usability" focuses primarily on the usefulness of evaluations, while active use plays only a minor role, and benefit is not examined.[5]

Figure 1 Assignment of the 37 quality criteria to the standard clusters and OECD-DAC criteria



Source: DEval, authors' own graphic

Note: blue bar = quality criterion examined by evaluation; yellow bar = quality criterion examined at the level of the organisation across all evaluations; org. = organisational, IR = Inception Report; ToC = Theory of Change; ToRs = Terms of References. [a] Quality criterion comes from the area of overlap between OECD-DAC and DeGEval standards; [b] Quality criterion comes from the "OECD-DAC only" area. The organisation-specific quality criteria and the quality criteria of the sustainability meta-evaluation are not shown, as they were not examined for all organisations.

3 Organisation-specific quality criteria were defined as requirements that are of great importance to an organisation for the quality of its evaluations, independently of the OECD-DAC or DeGEval standards.

4 The terms used to designate these three standard clusters display similarities to the designations of the DeGEval standards groupings (1. usefulness, 2. feasibility, 3. fairness and 4. accuracy). However, as the identified quality criteria partly represent the overlap between the OECD-DAC and the DeGEval standards, naming the clusters identically would not properly reflect their content. The OECD-DAC standards are structured largely in relation to evaluation phases, hence these terms were not used.

5 Details on the definitions can be found in section 2.1 of the main body of the evaluation report.

Factors affecting application of the quality criteria

The factors described below were analysed if they possessed three attributes, namely: 1) They had a clear cross-organisational definition. 2) It was possible to describe clear links to selected quality criteria. 3) Data were available either from the organisations or in secondary databases. To identify the factors, three focus group discussions were conducted with the responsible officers of the involved organisations, and scientific and empirical literature was reviewed. The factors were then systematically categorised as either 1) country-specific, 2) evaluation-specific or 3) organisation-specific.

Methodology

Data and data analysis

Once the organisations had been included in the meta-evaluation, a stratified random sample of a total of 296 evaluations was drawn for analysis of the application of quality standards. The organisations were selected using four criteria. This was in order to cover the greatest possible structural heterogeneity (criteria 1 to 3), and enable the analysis of a sufficient number of evaluations per organisation (criterion 4). In total, the evaluation units/desks of the organisations were (co-)responsible for 839 evaluations in Germany during the period under review from October 2016 to December 2020. The population then encompassed 576 evaluations that the BMZ had either (co-)funded, or in which a development intervention (co-)funded by the BMZ was examined. The sample drawn from that comprised 296 evaluations.

The understanding of quality (evaluation question 1): To determine the understanding of quality among the organisations involved, and which quality standards they were required to apply, the meta-evaluation examined documents of the organisations as well as relevant agreements with the BMZ. It also conducted a qualitative content analysis.

OECD-DAC and/or DeGEval standards (evaluation question 2a): To examine the application of the quality criteria in the 296 evaluations, evaluation documents (evaluation reports and annexes, ToRs, inception reports) and organisational documents (such as evaluation plans, guides and manuals for conducting evaluations) were used. In the inter-coding phase, for 14 of the 37 OECD-DAC and DeGEval quality criteria either very little information or none at all could be coded in the evaluation documents provided by the organisations. To avoid drawing erroneous conclusions about non-application, in a further step the responsible officers of the evaluation units/desks were therefore surveyed concerning the application of these quality criteria in their organisation.[6] The quality criteria described in the analysis grid were coded on the basis of ordinal or binary scores. Values were also assigned to the average frequencies of application of the quality criteria reported by the responsible officers of the evaluation units/desks in the online survey. For each quality criterion, different statistics (e.g. mean values) were calculated for each organisation. These statistics were subsequently converted into percentages and assigned to the predefined thresholds. A mean value across all organisations was then calculated, and analysed using descriptive statistical methods.

Organisation-specific quality standards (evaluation question 2b): To examine the application of the eleven organisation-specific quality criteria, these were coded in the evaluation documents of the four organisations concerned. Values were then calculated and analysed as with the OECD-DAC and the DeGEval quality criteria.

6 Since the quality criteria of the online survey were applied on average about 6 per cent less than the quality criteria of the document analysis, there was no reason to assume that the organisations systematically rated themselves more highly than they would have been rated by the objective coding.

Quality criteria of the sustainability meta-evaluation (evaluation question 2c): To analyse the application/repeated application of the quality criteria from the sustainability meta-evaluation, 15 quality criteria from the previous sustainability meta-evaluation of Noltze et al. (2018) were used. Eight quality criteria were already included as quality criteria in the OECD-DAC and in the DeGEval analysis grid and were transformed. The remaining seven were coded anew in the current evaluations. Descriptive statistics were calculated for the analysis. Structural equation models were calculated for the difference in findings between the sustainability meta-evaluation and the present one (Weiber and Mühlhaus, 2010).

To investigate links between selected factors and the application of the quality criteria, multivariate regression analyses were estimated (evaluation question 3). Regression analyses allow the identification of statistical correlations between the factors (independent variables) and the quality criteria as well as the standard cluster "reporting and methods" (dependent variables; Backhaus et al., 2011). The factors examined were either 1) country-specific, 2) evaluation-specific or 3) organisation-specific. The information for each factor was obtained from data submitted by the organisations or obtained from secondary databases.

Evaluating the application of quality standards

For organisations required to apply the quality criteria (group 1), application of the quality criteria was both analysed and rated. For organisations with no such requirement (group 2), application of the quality criteria was only analysed. The thresholds for rating the application of the quality criteria were defined in dialogue with the reference group. The threshold represented the ex ante assessment of when a quality criterion was considered to be barely applied, partly applied, largely applied or applied in an evaluation. When setting the thresholds in increments of 25 per cent (0 ≤ 25 per cent = "barely applied", > 25 ≤ 50 per cent = "partly applied", > 50 ≤ 75 per cent = "largely applied", > 75 ≤ 100 per cent = "applied"), the evaluation took into account the fact that the quality standards are to be understood as maximum standards. The rating was based on the defined thresholds for application, and by adding the extreme values 0 (not achieved) and 100 (exceeded) for group 1. Since application of the quality criteria of the sustainability meta-evaluation was now being analysed once again, at this point it was postulated that the application of the quality criteria had improved since that meta-evaluation. Both groups were analysed with regard to their degree of application of the quality criteria for the OECD-DAC and the DeGEval standards. Group 1 was also rated. For the OECD-DAC criteria (BMZ, 2006), the organisation-specific criteria and the quality criteria of the sustainability meta-evaluation, only organisations in group 1 were analysed and rated.

Strengths and challenges in the methodology

The selection of the organisations based on their structural heterogeneity enabled the meta-evaluation to analyse and describe the application of individual quality criteria across a correspondingly wide range. The findings of this meta-evaluation are valid for the involved organisations. The official implementing organisations are fully described. The selection of the involved non-governmental organisations is not representative for all non-governmental development organisations. The selection of organisations enables non-governmental organisations that were not involved to locate themselves within this range and use the findings of the meta-evaluation for themselves. The transferability of the findings to the population of evaluations by organisation was ensured by the selected statistical parameters for sampling. Transferability to other evaluation types of an organisation is not.

Since the analysis grid was derived from the OECD-DAC and DeGEval standards, it can also be used by other organisations. The analysis grid of this meta-evaluation can thus be used in the future to produce an analysis grid based on the BMZ Evaluation Policy.

When analysing the quality criteria from the online survey, there were limitations with regard to the triangulation of methods. Given the cost-benefit ratio, however, adding a further data collection method would not have been appropriate. Besides surveying the responsible officers of the evaluation units/desks, the triangulation of data might also have included the assessment of the evaluation teams for the respective evaluations. However, this meta-evaluation covered a large number of evaluations.

It was therefore beyond its scope to interview former evaluators from those evaluations instead of or in addition to the officers responsible for the evaluations. Furthermore, in some cases staff turnover would have made it no longer possible to locate and interview some of the evaluators. The sample size would thus have been reduced.

Generally speaking, there were limits to measuring some quality criteria. For certain quality criteria, it would take a lot of effort to investigate a "good" application. There are quality criteria that can only be examined in depth with a great deal of effort. For example, for the quality criteria "stakeholder involvement" and "accessibility for stakeholders", both the appropriate number of stakeholders who can be involved and the intensity of their involvement in the various evaluation phases are difficult to determine.

The operationalisation of the quality criteria was developed across all organisations. Some of these operationalisation choices did not match the evaluation practice of all organisations. This meant that for these organisations, application received a lower score than it would have if the criteria had been operationalised otherwise. There are quality criteria for which there was greater leeway (such as "quality assurance processes"). Accordingly, there is a conflict of objectives here between the meta-evaluation's interest in the application of selected quality criteria across organisations, and the heterogeneity of the application of the quality criteria.

Due to an inconsistent understanding of the measurement of "evaluation costs" across organisations, it was only possible to analyse explanations of the application/non-application of quality standards to a limited extent.

Examining the quality criteria of the sustainability meta-evaluation once again enabled the present meta-evaluation to analyse the difference in the application of the quality criteria by the GIZ and KfW over time. This provides evidence of the extent to which application of the quality criteria was improved through internal organisational reforms and the support of external actors (BMZ and DEval). It also brought to light challenges associated with a longitudinal study (such as raising the thresholds, and where necessary making appropriate adjustments to quality criteria over time).

Findings

Evaluation question 1: What understanding of evaluation quality do the involved German development cooperation organisations have?

The involved organisations' understanding of quality was predominantly based on the OECD-DAC and/or DeGEval standards and, where applicable, organisation-specific quality standards. When the meta-evaluation began, the involved organisations had in some cases not addressed these quality standards systematically. Furthermore, the BMZ's requirements concerning the application of the quality standards varied in the selected budget items during the period under review – in some cases the OECD-DAC standards were marked as mandatory, in others no requirements were specified.

Evaluation question 2a: To what extent are strengths and weaknesses evident in the application of the OECD-DAC and the DeGEval standards in the evaluations of the involved German development cooperation organisations?

Overall, a positive picture emerged regarding the application of the OECD-DAC and DeGEval standards. The involved German development cooperation organisations applied the quality standards in about two thirds of their evaluations. This was also the case – to a somewhat lesser degree – for organisations not required to apply the quality standards. Application of the quality standards sometimes varied widely between the organisations. This was to be expected, given the selection criteria for the inclusion of the involved organisations in the sample. It was thus possible to obtain a heterogeneous picture across the varying degrees of application.

However, it emerged that the organisations had largely not yet fully identified the quality standards in their organisational documents. Nor had they systematically prescribed the application/ non-application of these quality standards. This was also the case for the traceability of the application or

non-application of some selected quality standards at the level of the individual evaluation. It should be noted that, for various reasons, the application of some quality standards was recorded not at the evaluation level, but at the organisational level. This might be due to either the way in which the meta-evaluation chose to operationalise some quality criteria, or a lack of documentation on application/ non-application, or the fact that application was documented exclusively at the organisational level rather than at the level of the individual evaluation. There is a clear need for improvement here, as without information, an external investigation of an (explained) application/non-application of the quality standards at the level of the evaluation is only possible to a limited extent. It was thus not possible to trace whether a quality standard was either not applied (with or without explanation), or was applied but was not documented.

The OECD-DAC criteria were very largely achieved by the involved German development cooperation organisations. However, it should be explicitly pointed out that operationalisation took place in relation to the OECD-DAC standards rather than the OECD-DAC criteria. This made it easy to achieve the threshold. In the application of the OECD-DAC criteria, first documentation of non-application was also available at the organisational and evaluation levels. In this respect, the application of the OECD-DAC criteria already differed – albeit to a minor extent – from the application of most of the other quality criteria. It can be assumed that in future evaluations, documentation of the non-application of the OECD-DAC criteria will continue to increase. This is because since 2020 (BMZ, 2020), the updated BMZ guidelines on evaluation criteria require priority setting that is explained and transparent.

In the annex to the report in section 7.1, the findings for the four official implementing organisations BGR, GIZ, KfW and PTB are also presented and classified at the level of the individual organisation.

Evaluation question 2b: To what extent are strengths and weaknesses evident in the application of the organisation-specific quality standards in the evaluations of the involved German development cooperation organisations?

A positive picture also emerged for the application of the organisation-specific quality standards by DRK, EWDE, GIZ and hbs. On average, these quality criteria were "largely achieved". Once again, there was potential for improvement in the explanation of non-application at the evaluation level.

Evaluation question 2c: To what extent are strengths and weaknesses evident in the application of the sustainability meta-evaluation quality criteria in the evaluations of the GIZ and KfW?

Regarding the application of these quality criteria, the picture was a positive one – with a few exceptions. Specifically, the quality criteria were achieved on average to a degree of about 75 per cent. This represents a somewhat higher degree of application than was found for the OECD-DAC and DeGEval standards. However, challenges remained in the application of the quality criteria "selection procedure for interviewees described" and "control/comparison groups included". Furthermore, it should be noted that the application of all the quality criteria has improved – in some cases clearly – since the sustainability meta-evaluation. Overall, an average difference of 36 per cent was observed. These changes indicate that the measures implemented after the sustainability meta-evaluation to improve the evaluation practices of GIZ and KfW might have affected application. This is a very positive result in light of the extensive efforts made by a large number of actors in connection with the measures. It should be noted, however, that alternative explanations cannot be ruled out (for example, operationalisation of the quality criteria in ways that make it relatively easy to achieve the thresholds, or changed documentation methods).

Evaluation question 3: To what extent are country-specific, evaluation-specific and organisation-specific factors linked to the application of quality standards?

Overall, the findings provide little evidence of significant links between the factors identified in the literature and the focus group discussions, and the application of the quality criteria. This involves mainly evaluation-specific factors: “number of internal and external evaluators“ and the implementation of various “quality assurance processes“.

Conclusions and recommendations

The meta-evaluation's recommendations are derived primarily from the evaluation questions on the application of selected quality standards and quality criteria (OECD-DAC, DeGEval and/or organisation-specific quality standards, as well as quality criteria of the sustainability meta-evaluation, evaluation questions 2a to 2c). The recommendations are worded in general terms. This means that each involved organisation must consider its own organisation-specific findings in order to determine the particular relevance of the recommendations to it. This is because the cross-organisational mean value presented in the findings is not sufficiently meaningful for evaluating the application of quality standards by specific organisations. The criteria-based selection of non-governmental organisations focuses on their structural heterogeneity, and thus reflects the range of possible degrees and forms of application for different organisations. This means that non-governmental organisations which were not involved can also determine the relevance of the findings to them, and thus draw guidance from the conclusions and recommendations. The recommendations addressed to the BMZ are intended for the BMZ Evaluation Division.

Identification of relevant/non-relevant quality standards and prescription thereof in organisational documents

Recommendation 1

a) As part of a revision of their evaluation practice, the evaluation units/desks of the BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR and PTB should (if they have not already done so) identify the quality standards that are required for their organisation. They should explicitly prescribe these in organisational documents, and define their application in evaluation processes. The organisations should review at regular intervals the identification and systematic prescription of quality standards. When doing so, they should specify the degree of application they require for each of the quality standards.

b) In the context of upcoming updates of funding guidelines or ancillary provisions for individual budget items, the BMZ should make a contribution towards strengthening its Evaluation Policy as a reference document for evaluations. As part of these updates, together with the non-governmental organisations concerned the BMZ should establish and prescribe special conditions for particular organisations (e.g. as in the case of the funding guidelines for political foundations). The principle of maximum standards should be retained here.

c) Based on its Evaluation Policy, and in dialogue with the official implementing and non-governmental organisations, the BMZ should develop an analysis grid for the application of the quality standards, taking into account the analysis grid of the present meta-evaluation. It should also make this available to the official implementing and non-governmental organisations.

Recommendation 2

a) The evaluation units/desks of the BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR and PTB should explain and prescribe in organisational documents any general non-application of particular quality standards that are required for them.

b) The BMZ should reach an agreement with the official implementing organisations on the application and (explained) non-application of the quality standards described in the BMZ Evaluation Policy. It should do so either in order to jointly determine non-application at the organisational level or to document discrepancies.

Ensuring the application and traceability of the application/non-application of relevant quality standards at evaluation level

Recommendation 3

a) The evaluation units/desks of the BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR and PTB should, if they have not already done so, further improve the application of the quality standards required at the organisational level (recommendation 1) in individual evaluations, and especially those quality standards that are barely or party applied. Furthermore, the application or (explained) non-application of all quality standards should be traceable at the level of each evaluation and regularly reviewed by the organisations.

b) The BMZ should urge the official implementing organisations to ensure the application of the relevant quality standards, and the traceability of their application/non-application, at evaluation level.

Joint learning

Recommendation 4

a) The evaluation units/desks of the BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR and PTB, and representatives of VENRO, should regularly share their various lessons learned in identifying, prescribing, assuring and tracing the application/non-application of all quality standards. This dialogue shoswuld also integrate non-involved organisations and include further types of evaluation – such as decentralised evaluations – in order to continue improving the application of quality standards.

b) The BMZ should financially support the dialogue with and between the organisations on identifying, prescribing, assuring and tracing the application/non-application of the quality standards.

Ensuring the application and traceability of the application/non-application of the sustainability meta-evaluation quality criteria

Recommendation 5

a) When developing the analysis grid for the quality standards described in the BMZ Evaluation Policy (recommendation 1), the BMZ should consider adopting the quality criteria from the sustainability meta-evaluation. If appropriate, it should also include them in the analysis grid.

b) Based on recommendation 5a, the GIZ and KfW should ensure/improve the application/non-application of the quality criteria from the sustainability meta-evaluation that have been incorporated into a BMZ analysis grid. They should also guarantee the traceability of the (explained) application/non-application for each evaluation.

CONTENTS

Figures

Tables

Boxes

ABBREVIATIONS AND ACRONYMS

BGR	German Federal Institute for Geosciences and Natural Resources
BMZ	German Federal Ministry for Economic Cooperation and Development
CARE	CARE Germany
DAC	Development Assistance Committee
DeGEval	The Evaluation Society in Germany
DRK	German Red Cross
DVV	Institute for International Cooperation of the German Adult Education Association
EWDE	Protestant Agency for Diakonie and Development
FTE	Full-time equivalent
GIZ	Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH
hbs	Heinrich Böll Foundation
IR	Inception report
KAS	Konrad Adenauer Foundation
KfW	KfW Development Bank
OECD	Organisation for Economic Co-operation and Development
PTB	German National Metrology Institute
ToR	Terms of Reference

1. INTRODUCTION

This introductory chapter covers the background, the evaluation object, the involved organisations, and the objectives and evaluation questions of the meta-evaluation.

1.1 Background

Evaluations are a recognised means of examining how successful a development intervention has been. They generate new evidence for decision-makers, and promote learning from conclusions and recommendations. They also strengthen accountability by providing independent assessment. To achieve this, evaluations "[...] provide the most objective and empirically sound analyses and assessments possible of the extent to which development interventions have achieved results" (BMZ, 2021a: 4). The target groups of evaluations often comprise those responsible for the development intervention and their partners, as well as policy makers. However, they also include for instance the evaluation community, and the target groups of the evaluated development intervention itself. To ensure that evaluations are of high quality, quality standards should be applied when carrying them out.[7] Applying these quality standards ensures that key aspects are taken into account in the design, implementation and use of an evaluation. (Examples include an analysis of the development intervention that is as robust as possible, the participation of various stakeholders and the usefulness of the recommendations). To independently examine the application of such standards across selected German development cooperation organisations, the "Meta-evaluation on the quality of (project) evaluations in German development cooperation" was included in DEval's multi-year evaluation programme (DEval, 2021b).

Meta-evaluations can be described as evaluations of evaluations. They are becoming increasingly important in development cooperation. According to Caracelli and Cooksy (2009: 2 f.), meta-evaluations are conducted "to improve an evaluation in process, reflect systematically on the strengths and weaknesses of an evaluation to enhance one's future evaluation practice, or provide information about the credibility of the findings to users."[8] This understanding is also adopted in the present meta-evaluation. Meta-evaluations differ from meta-analyses and evaluation syntheses. The former summarise and analyse various findings quantitatively The latter perform an evaluation of evaluations with a thematic focus (Caspari, 2012). There are now a large number of meta-evaluations in international development cooperation. (Examples include those conducted by the Austrian Development Agency and the Ministry for Foreign Affairs of Finland.) In German development cooperation too, meta-evaluations are now being carried out within organisations (such as the Deutsche Gesellschaft für Internationale Zusammenarbeit [GIZ] GmbH, Welthungerhilfe, the Friedrich Ebert Foundation, MISEREOR and World Vision). Furthermore, DEval has conducted the first cross-organisational meta-evaluations (such as the meta-evaluation of the project evaluations of GIZ and KfW Development Bank [KfW]).[9]

This meta-evaluation ran concurrently with the preparation of the Evaluation Policy for German Development Cooperation, published by the Federal Ministry for Economic Cooperation and Development (BMZ). It follows on from previous assessments of the German development cooperation system, the study to monitor implementation of the system assessment, and the sustainability meta-evaluation (Noltze et al., 2018), among others. In 1999, the BMZ commissioned a first systematic assessment of evaluation practice in German development cooperation (Borrmann et al., 1999). Ten years later, a second system assessment was completed. The latter contains system-wide recommendations for the further development of evaluation practice (Borrmann and Stockmann, 2009). As part of this, guidelines were developed that took

7 The OECD (2022) also currently recommends promoting the quality of evaluations in the area of public policy, for example by setting standards.

8 Another definition of meta-evaluation by the OECD-DAC (2002: 27) states: "The term is used [...] to denote the assessment of an evaluation to judge its quality."

9 Examples of other meta-evaluations in which quality is examined in relation to various internationally recognised or specially developed quality standards include Caspari (2010, 2011), FES (2015), Freiman et al. (2016, 2017), Hageboeck et al. (2013), HTSPE Limited (2011), Koy et al. (2016), Krämer et al. (2019), Mauthofer and Silvestrini (2018), Noltze et al.(2018), Queiroz de Souza (2017), Silvestrini and Bäthge (2019), Silvestrini et al. (2018), UNFPA (2020) and Väth et al. (2022). A list of meta-evaluations carried out can be found in section 1 of the online annex.

into account, among other things, a preliminary version of the standards of the Development Assistance Committee (DAC) of the Organisation for Economic Co-operation and Development (OECD), the first version of the standards of the Evaluation Society in Germany (DeGEval), the OECD-DAC Criteria and the OECD-DAC Principles. Building on this, in 2015 DEval conducted a study to monitor implementation of the findings and recommendations of the 2009 system assessment (Lücking et al., 2015).[10] Finally, in 2018, DEval published a cross-organisational meta-evaluation[11] of the quality of GIZ and KfW project evaluations (Noltze et al., 2018). Both that meta-evaluation and the present one examine the quality of those evaluations. They do not examine the quality of the evaluation system in German development cooperation.

During the above-mentioned period, there were regular exchanges between various evaluation units/desks of German official implementing (governmental) and non-governmental organisations and the BMZ on the quality of evaluation practice. The BMZ published its Evaluation Policy for German Development Cooperation (BMZ, 2021a) prior to the completion of this meta-evaluation, thus setting the most recent milestone. This policy further consolidates the understanding of evaluation quality in German development cooperation. In particular, the policy states that both the OECD-DAC standards and the DeGEval standards are binding for the official implementing organisations, and provide guidance for the non-governmental organisations. Since the period of the meta-evaluation preceded the adoption of the guidelines, it was not possible to include them in the study. Nevertheless, important findings were generated for the future unfolding of the policy.

The meta-evaluation included evaluations from eleven German official implementing and non-governmental development cooperation organisations. This permits a cross-organisational view of the application of quality standards in German development cooperation. The meta-evaluation examined project evaluations for which the evaluation units/desks of German organisations were (co-)responsible and which were (co-)funded by the BMZ. All these evaluations were implemented between October 2016 and December 2020[12] **.** BMZ funding was in place if either the project evaluation or the development intervention of that evaluation was (co-)funded. Project evaluations[13] were included, but strategic evaluations (e.g. meta-evaluations or corporate strategic evaluations) and decentralised evaluations were not.[14] One particular feature of the present meta-evaluation is that it examines the application of quality standards in evaluations for a variety of official implementing organisations (the Federal Institute for Geosciences and Natural Resources [BGR], GIZ, KfW, the National Metrology Institute [PTB]), and non-governmental organisations (CARE Germany [CARE], the German Red Cross [DRK], the Institute for International Cooperation of the German Adult Education Association [DVV], the Protestant Agency for Diakonie and Development [EWDE], the Heinrich Böll Foundation [hbs], the Konrad Adenauer Foundation [KAS] and MISEREOR). At the time of the meta-evaluation, DEval's mandate to evaluate development cooperation actors was clearly defined by the BMZ Guidelines for Bilateral Financial and Technical Cooperation only for the four official implementing organisations. Nonetheless, the other organisations also accepted the offer of being examined as part of an independent, external review.

10 Since the system assessment and the study to monitor it also made reference to the OECD-DAC and DeGEval standards, there are points where they could be cross-linked with the present meta-evaluation. These are discussed in further detail in section 4.1.4 of the online annex, but are not the focus of this meta-evaluation.

11 The BMZ Evaluation Policy states that DEval meta-evaluations are part of the quality assurance framework for the evaluation system (BMZ, 2021a).

12 The time period was defined so as to ensure comparability with the sustainability meta-evaluation (Noltze et al., 2018). Evaluations are available from the Heinrich Böll Foundation from January 2016 to October 2020, and from CARE and GIZ from January 2018 to December 2020.

13 Henceforth, "project" is used only rarely as a prefix, as not all organisations evaluate single development interventions. They sometimes also evaluate country offices.

14 There are also "bundled evaluations", which are considered a hybrid form between project and strategic evaluations. In some cases, these were added to the population (for example, in the case of a more technical bundle) or omitted (for example, in the case of a more strategic issue). The terms described are imprecise, as project evaluations can also be used strategically. However, they are in line with the evaluation practice of a large number of organisations and are therefore retained here.

When selecting the organisations, attention was paid to their structural heterogeneity so that the findings could reflect the broadest possible range of experience with the application of quality standards. The heterogeneity of the organisations was reflected, among other things, in the size of the evaluation units/desks: The average number of full-time equivalents (FTEs) in the evaluation units/desks per year during the period under review ranged from 0.5 to 20. The size of the evaluation units/desks relative to the size of the organisation (number of FTEs per year for the evaluation units/desks relative to the number of FTEs for the organisation) also differed between the organisations. It ranged from 0.2 to 5.2 per cent. The approximate average BMZ funding of development interventions (i.e. the value of commissions) per year varied between 1.3 million and 1.5 billion euros. The number of evaluations for which the evaluation units/desks were co-responsible ranged from two to 57 per year. This focus on the heterogeneity of the organisations made it possible to examine different structural challenges in the application of quality standards, and take these into account in the conclusions and recommendations. This means that the conclusions and recommendations are also potentially relevant to organisations that were not involved. Heterogeneity exists not only between the organisations, but also between the evaluations. For example, the evaluations were implemented in different regions (Europe, Africa, Asia and the Americas) and in different sectors (such as social infrastructure, productive sectors, indebtedness, humanitarian aid, climate and gender).

The application of quality standards was examined in relation to whether or not the organisations are required to apply the standards documents[15] of OECD-DAC and/or DeGEval, as well as organisation-specific standards. Furthermore, the findings for the official implementing organisations are treated separately, as these organisations are required to implement the BMZ guidelines indirectly. For non-governmental organisations the guidelines provide guidance. The application of quality standards was considered required if, during the period under review, written instructions to apply them were (i) included in the involved organisation's own documents, or (ii) included as part of the BMZ Guidelines for bilateral Financial and Technical Cooperation (BMZ, 2021b), or (iii) included in other binding documents for the involved organisations (e.g. funding guidelines of the "private German organisations" or "agencies engaged in social improvement"), or if (iv) the organisation was a member of DeGEval. Accordingly, the organisations were divided into two groups for the study: Group 1 comprised involved organisations with a requirement to apply quality standards. Group 2 comprised involved organisations with no such requirement. Furthermore, a distinction is drawn between official implementing (governmental) and non-governmental organisations that also has a history. The term "non-governmental organisations" refers to engaged civil society actors, municipalities and private-sector actors. Based on the primacy of "different instruments with different roles" in bilateral official development cooperation, governmental organisations must implement the BMZ Guidelines indirectly. For non-governmental organisations, on the other hand, the Guidelines provide guidance. Accordingly, in the annex to this report (section 7.1) the findings for official implementing organisations are shown separately. Furthermore, in the recommendations drawn up for BMZ, distinctions have been made in some cases between official implementing and non-governmental organisations.

1.2 Objectives of the evaluation and evaluation questions

This cross-organisational meta-evaluation aims to generate findings on the involved organisations' understanding of the quality of evaluations, and on strengths and weaknesses in the application of quality standards. It also identifies and analyses factors that may explain the degree to which quality standards are applied. A first objective was to determine the understanding of quality in evaluations that is present among different German development cooperation organisations (learning). The main objective was to identify whether and, if so, to what extent the organisations involved applied the two internationally

15 The "Quality standards for development evaluation" (OECD-DAC, 2010) and the "*Standards für Evaluation*" (DeGEval, 2016) are hereinafter referred to as "standards documents".

recognised quality standards for evaluations – the OECD-DAC and DeGEval standards[16] – and/or organisation-specific quality standards. Based on the involved organisations' understanding of quality, a further aim was to examine their application of quality standards (learning and accountability). Furthermore, this meta-evaluation analysed the extent to which GIZ and KfW have improved their application of the quality criteria in their evaluation practice following the sustainability meta-evaluation (Noltze et al., 2018). Finally, to support future learning, the meta-evaluation also aimed to find explanations for the non-application of quality standards (learning).

Evaluation question 1: What understanding of evaluation quality do the involved German development cooperation organisations have?

During the period under review, the involved organisations had a heterogeneous understanding of evaluation quality. In a first step, it was therefore necessary to capture, systematically analyse and present this. In doing so, a distinction was drawn between internationally recognised and organisation-specific quality standards.

Evaluation question 2: To what extent are quality standards applied in evaluations of the involved German development cooperation organisations?

To answer this question, the quality standards were divided up into three areas and examined accordingly. These were: 1) internationally recognised quality standards, meaning the OECD-DAC and DeGEval standards; 2) organisation-specific quality standards, and 3) quality criteria of the sustainability meta-evaluation. A separate evaluation question was formulated for each area.

Evaluation question 2a: To what extent are strengths and weaknesses evident in the application of the OECD-DAC and the DeGEval standards in the evaluations of the involved German development cooperation organisations?

Internationally recognised quality standards describe attributes for "quality evaluations".[17] To be able to examine the involved organisations with a coherent analysis grid, the team identified the overlap between the two relevant international sets of quality standards – OECD-DAC standards as quality standards for evaluations in development cooperation in particular, and DeGEval standards as quality standards for evaluations in general. Individual OECD-DAC standards not included therein were also analysed (normative frame of reference)

Evaluation question 2b: To what extent are strengths and weaknesses evident in the application of the organisation-specific quality standards in the evaluations of the involved German development cooperation organisations?

There are also organisation-specific quality standards that go beyond the quality standards of the OECD-DAC and DeGEval, and are of particular importance to the organisations. These are reflected accordingly in the respective evaluations, and were also examined.

16 Other international standards documents include e.g. the African Evaluation Guidelines of the African Evaluation Association (AfrEA, 2020), the Evaluation Standards for Latin America and the Caribbean (Rodríguez Bilella et al., 2016) and the Norms for Standards and Evaluation of the United Nations Evaluation Group (UNEG, 2016). The BMZ Evaluation Policy was not yet in force at the time of the analysis, hence it was not possible to consider it in the present study.

17 OECD-DAC (2010: 5) describes this aspiration in its standards document as follows: "The DAC Quality Standards for Development Evaluation identify the key pillars needed for a quality development evaluation process and product". DeGEval also emphasises the importance of its quality standards. The quality standards "should give evaluators as well as commissioning parties […] guidance on how to design good evaluations" (DeGEval, 2016: 5). The current BMZ Evaluation Policy (2021a) states that the Ministry considers both the OECD-DAC standards and the DeGEval standards to be relevant (the OECD-DAC standards particularly so).

Evaluation question 2c: To what extent are strengths and weaknesses evident in the application of the sustainability meta-evaluation quality criteria in the evaluations of the GIZ and KfW?

Furthermore – based on the study and the findings of the sustainability meta-evaluation (Noltze et al., 2018) – the application of the additional quality criteria was examined again for GIZ and KfW. These findings were then compared with the previous ones.

Evaluation question 3: To what extent are country-specific, evaluation-specific and organisation-specific factors linked to the application of quality standards?

To enable those responsible within the involved organisations to learn more about the application of quality standards, factors affecting it were identified and empirically analysed. To ensure that these factors were investigated systematically, they were each defined as either 1) country-specific, 2) evaluation-specific, or 3) organisation-specific.

2. THEORETICAL AND EMPIRICAL BACKGROUND

In this chapter, the first section describes the understanding of quality in this meta-evaluation. It then explains the analysis grid with the quality criteria used, and the allocation of these quality criteria to the three standard clusters "reporting and methods", "participation, independence and fairness" and "usability". The second section identifies relevant factors that may be positively or negatively linked to the application of quality standards.

2.1 Understanding of quality, analysis grid and standard clusters

To examine the application of quality standards at the involved organisations, an agreed understanding of quality as well as an analysis grid that is relevant to all involved organisations is required. Therefore, the analysis grid was systematically derived from the OECD-DAC and DeGEval standards documents, the organisations' own documents and the quality criteria of the sustainability meta-evaluation. As the BMZ Evaluation Policy was being prepared at the same time as the analysis grid for this meta-evaluation, it was not possible to take the policy fully into account.

The understanding of quality

This meta-evaluation equates evaluation quality with the application of the relevant quality standards, i.e. the quality standards that are required for the involved organisations. It then examines evaluation quality accordingly. The term "application of quality standards" was chosen in consensus with the reference group[18]. It describes the extent to which evidence can be obtained that quality standards were addressed in the evaluations examined, and that the application of those standards was ensured. (In other words, either this was documented in writing, or was reported back in writing upon request by the evaluation team.) The standards of the OECD-DAC and DeGEval are fundamental for the quality of evaluations, due to their international recognition[19], their link to development cooperation and their relevance to German development cooperation organisations.[20] These internationally recognised quality standards were complemented with organisation-specific quality standards and the quality criteria of the sustainability meta-evaluation. Since other understandings of quality do exist, a high degree of "application of quality standards" in evaluations does not mean that this would also receive a high rating given an alternative understanding of quality.[21]

The OECD-DAC and DeGEval standards are maximum standards. This means that the involved organisations do not have to apply all quality standards in all evaluations. For various reasons, the involved organisations might not apply particular quality standards. This may be the case for instance because (i) those standards do not match the (conceptual) orientation of an organisation's evaluation strategy (e.g. "consideration of joint evaluations" in the case of non-governmental organisations), (ii) they cannot be applied in some evaluations (e.g. ensuring "publication of the evaluation report", for reasons of confidentiality), or (iii) they are mutually exclusive (where applicable, "description of the methodological adequacy" and "timeliness of the findings"). Furthermore, the application of particular quality standards can vary greatly between organisations. This is because these organisations operate in different countries, sectors and contexts, or have different objectives and values that can influence which standards they apply.

18 The reference group comprised representatives of the involved organisations and VENRO, and officers from BMZ Division GS 22 "Evaluation and development research, DEval, IDOS". The members accompanied the evaluation process during all phases of the evaluation (for example through virtual meetings or comments on evaluation documents; DEval, 2021a).

19 A first draft of the OECD-DAC standards was piloted between 2006 and 2009. This pilot was revised in response to the comments of the members, and introduced in 2010 (OECD, 2013). The OECD-DAC standards document states: "Built through international consensus, the standards are intended to serve as an incentive and inspiration to improve evaluation practice" (OECD DAC, 2010: 1).

20 The OECD-DAC and DeGEval standards are made binding for the official implementing organisations through the BMZ Evaluation Policy. For German civil society organisations they offer guidance (BMZ, 2021a).

21 In this meta-evaluation, for instance, an incomplete, non-random survey within an evaluation can be considered "applied" if the methodological adequacy is adequately described and the limitations are described (for example, in certain situations in fragile contexts). However, if the methodological approach were to be viewed from the perspective of rigorous impact evaluation, it could be judged to have failed.

In the present meta-evaluation, a maximum consideration of all quality standards was selected, in order to obtain a first, comprehensive picture of the current application practice of the involved organisations. However, the principle of maximum standards was taken into account for the assessment.

Analysis grid

The analysis grid contains 37 quality criteria derived from the OECD-DAC and DeGEval standards documents, eleven organisation-specific quality criteria and a further eight quality criteria from the sustainability meta-evaluation. [22] *The coding guide for the OECD-DAC and DeGEval quality criteria can be found in section 6 of the online annex.*

The analysis grid includes quality criteria derived from the OECD-DAC and DeGEval standards documents. These can be classified into three areas: 1) the overlap between the standards documents of OECD-DAC and DeGEval, 2) the OECD-DAC standards document minus the overlap with the DeGEval standards document ("OECD-DAC only") and 3) the OECD-DAC criteria. According to DeGEval (2016: 25), the standards are intended to "serve as criteria for meta-evaluation, i.e. the evaluation of evaluations, by defining what attributes good evaluations should display". The two internationally recognised standards documents combine some similar elements (such as "description of the evaluation object'" and some different elements (such as "partner-country orientation") in their understanding of quality. In the course of identifying the overlap between the two standards documents, it became apparent that passages from all DeGEval standards match passages of the OECD-DAC standards. In a next step, the passages that constitute the overlap were identified and assigned a quality criterion. Furthermore, OECD-DAC quality standards that do not overlap with the DeGEval standards ("OECD- DAC only") were also included in the analysis grid and assigned quality criteria (Table 1; an overview of the derivation of all OECD-DAC and DeGEval quality criteria can be found in section 2.1 of the online annex). [23] Seven quality standards were analysed using more than one quality criterion. Another area is the OECD-DAC criteria (Figure 2; DEval, 2020). Since all involved organisations were committed to the OECD-DAC criteria, these are listed as a separate area, although their application is part of the OECD-DAC standards (quality standard 2.8; OECD DAC, 2010).

A total of 26 quality standards were included in the analysis grid, to which 37 quality criteria are assigned. Where possible, during operationalisation the quality criteria were aligned with the criteria of the sustainability meta-evaluation (Noltze et al., 2018) or the study on monitoring of the system assessment (Lücking et al., 2015). The criteria represent both the operationalisation of the quality standards and the basis for rating.

[22] The term "quality standards" refers to the original text from the standards documents of the OECD-DAC and DeGEval. The "quality criteria" are derived from the quality standards for the purpose of this study. Some quality standards contain several elements. Accordingly, these were examined using several quality criteria.

[23] Concerning operationalisation of the standards, two things should be noted: 1) "Sub-elements" of the original quality standards were not examined, as these were not a significant part of the overlap (for example, OECD-DAC Standard 3.15 specifies that differences of opinion or interpretation in the evaluation report should be reproduced in footnotes or annexes. This specification did not represent a substantial overlap with DeGEval Standard N5, which is aimed at the transparent documentation of different perspectives, and does not describe whether this should take place in the footnotes or the annex). 2) Some terms from the standards documents with complex implications (e.g. "security") either were not broken down any further for operationalisation, or a specification chosen by the evaluation team was used. In both cases, an understanding that is consistent with the quality standards of the standard documents is not guaranteed. Furthermore, in the standards documents, the OECD-DAC standards are usually rather detailed and the DeGEval standards rather general. Consequently, the act of combining these passages with each other leaves room for interpretation of the text elements.

Table 1 Derivation of the quality criterion "description of the evaluation object"

Name	Standards document	Overlap	Quality criterion
Description of the evaluation object	OECD-DAC 2.3	"The development evaluation being evaluated (the evaluation object) is clearly defined" (OECD DAC, 2010: 8).	The quality criterion is a when (1) objective(s), (2) target groups(s) and (3) relevant actors (political partner and/or executing agency) of the development intervention are specified.
	DeGEval G1	"The conception of the evaluation object […] is described and documented precisely and comprehensively" (DeGEval, 2016: 20).	

Source: DEval, authors' own table

Note: The quality criterion was coded using a four-point rating scale. If none of the three items described in the quality criterion were described in the evaluation, it was rated as "not achieved" (1). If one was described, it was rated as "partly not achieved" (2), if two, as "partly achieved" (3), and if three, as "achieved"(4). Section 6 of the online annex presents the coding guide, which shows all the ratings for each quality criterion.

For DRK, EWDE, GIZ and hbs, the analysis grid included further organisation-specific quality criteria. Organisation-specific quality criteria constitute a further area. These were defined as requirements that are of great importance to an organisation for the quality of its evaluations, independently of the OECD-DAC or DeGEval standards. On this matter the OECD-DAC (2010: 5) states: "[...] these standards do not exclude the use of other evaluation quality standards and related texts, such as those developed by individual agencies, professional evaluation societies and networks." First, eleven organisation-specific quality criteria were identified from each organisation's documents and then operationalised in dialogue with that organisation. Besides thematic content, they also include various methods and the role of the partner in the development intervention. Content-related quality criteria of two organisations mainly refer to the inclusion of gender issues in various evaluation documents such as the Terms of Reference, and in the findings in the evaluation report. For the GIZ, contribution and efficiency analysis in the area of methods were identified as organisation-specific quality criteria. Another organisation focused on the role of the partner, for instance in the addressee orientation of the recommendations.

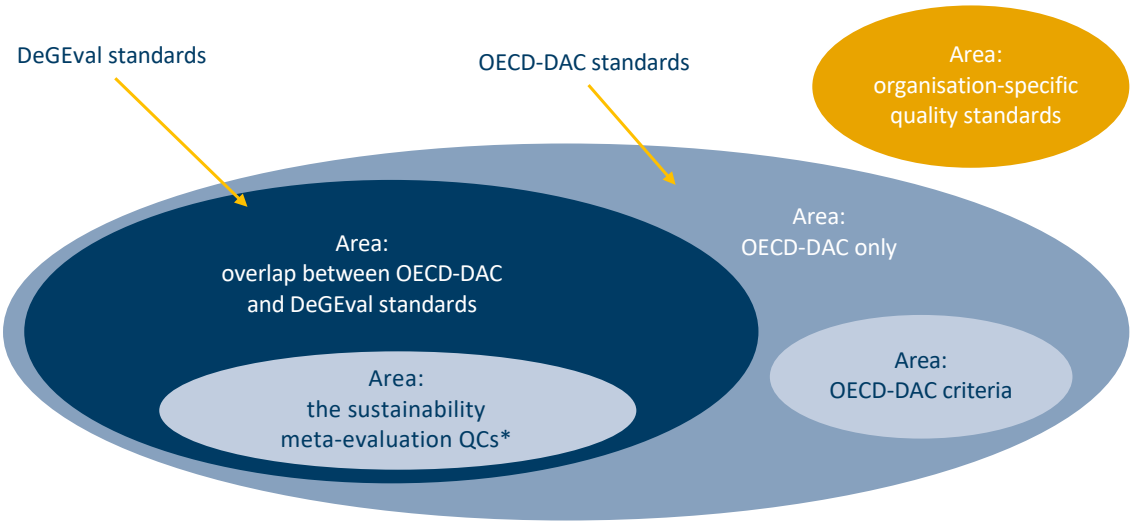
For KfW and GIZ, quality criteria already identified in the previous sustainability meta-evaluation were also included in the analysis grid and examined again. In terms of content, these quality criteria focus on the area of methodology (for example, the performance of a "before-and-after comparison"). Seven of the 15 quality criteria in this area were encoded once again in the present meta-evaluation. Here, some adjustments had to be made to the coding rules, as the structure and content of the evaluation reports had changed since the sustainability meta-evaluation.[24] The other eight quality criteria could each be assigned to a quality criterion from the analysis grid for the internationally recognised quality standards. The comparability of these eight (transformed) quality criteria is in some cases more limited than the comparability of those which were re-encoded.[25] The additional quality criteria were only examined for GIZ and KfW and not the other

[24] The quality criteria cannot be derived in full from the overlaps between the two standards documents, but can be assigned to them in terms of content. This concerns the quality criteria "indicators used", "selection procedure described", "before-and-after comparison", "control/comparison groups", "causality based on plausibility", "triangulation of methods" and "data basis adequate".

[25] This applies in particular to the quality criteria "methodology described" and "interviewees identified", as well as "conclusions referenced" and "conclusions plausible", as these were each compared with the same quality criterion of the present meta-evaluation. The quality criterion "triangulation of data" was not examined again, as its content was covered by the quality criterion "triangulation of methods". Further details can be found in section 2.1 of the online annex. In addition, two of the data collection methods not assessed in the sustainability meta-evaluation were examined (online annex, section 4.1.2).

nine involved organisations, as only evaluations of these two organisations had been examined in the sustainability meta-evaluation. Further details on the quality criteria of the sustainability meta-evaluation can be found in section 2.1 of the online annex.

Figure 2 The five areas of analysis



Source: DEval, authors' own graphic

*Note: * The quality criteria of the sustainability meta-evaluation cannot be derived in full from the overlap between the two standards documents, but can be assigned to them in terms of content.*

Standard clusters

With the exception of the five OECD-DAC criteria, the quality criteria were assigned to three standard clusters – "reporting and methods", "participation, independence and fairness" and "usability"[26]**.**[27] Since a quality criterion could be assigned to several standard clusters based on content (for example, "stakeholder involvement" can be both useful, as stakeholders may be more likely to take the evaluation findings into account, and fair, as they are then able to contribute to the evaluation), the criteria were assigned according to where the content was clearest in suggesting this (Table 2). The terms used to designate these three standard clusters display similarities to the designations of the DeGEval standards groupings (1. usefulness, 2. feasibility, 3. fairness and 4. accuracy). However, as the identified quality criteria partly represent the overlap between the OECD-DAC and the DeGEval standards, naming the clusters identically would not properly reflect the content.[28]

26 The name "usability" was chosen to avoid confusion with the "usefulness" cluster of the DeGEval standards. The "usability" cluster partly deviates from that and goes beyond it in terms of content.

27 The quality standards of the DeGEval "feasibility" cluster were not omitted, but were covered by various quality criteria in the standard clusters "participation, independence and fairness", and "usability".

28 The OECD-DAC standards are structured largely in relation to evaluation phases. The evaluation phases of the OECD-DAC standards are: 1. overarching considerations, 2. purpose, planning and design, 3. implementation and reporting, 4. follow-up, use and learning. The foreword of the DeGEval standards states that standards can be assigned to several evaluation phases, and that a chronological classification is therefore avoided: "Also, a restructuring based on typical phases of the evaluation process was deliberately avoided [,] as a restructuring of this kind would not be uniformly applicable to all evaluation cases" (DeGEval, 2016: 14). Nonetheless, for general orientation purposes chapter 6 of the DeGEval standards does assign the standards to six phases of an evaluation.

Table 2 Number of quality criteria per area and standard cluster

Area	Organisations examined	Reporting and methods	Participation, independence and fairness	Usability	Own cluster	Total
Overlap between the OECD-DAC and DeGEval standards	all	10	6	8	/	37
OECD-DAC only	all	1	5	2	/	
OECD-DAC criteria	all	/	/	/	5	
Total	***all***	***11***	***11***	***10***	***5***	***37***
Organisation-specific QCs	DRK, EWDE, GIZ, hbs	3	5	3	/	11
QCs sustainability meta-evaluation	GIZ, KfW	15	/	/	/	15

Source: DEval, authors' own table

Note: QC = quality criterion. 19 quality standards (QSs) were calculated from one quality criterion (QC), five QSs from two QCs, one QS from three QCs and one more QS from five QCs; accordingly, 37 quality criteria were obtained from 26 quality standards.

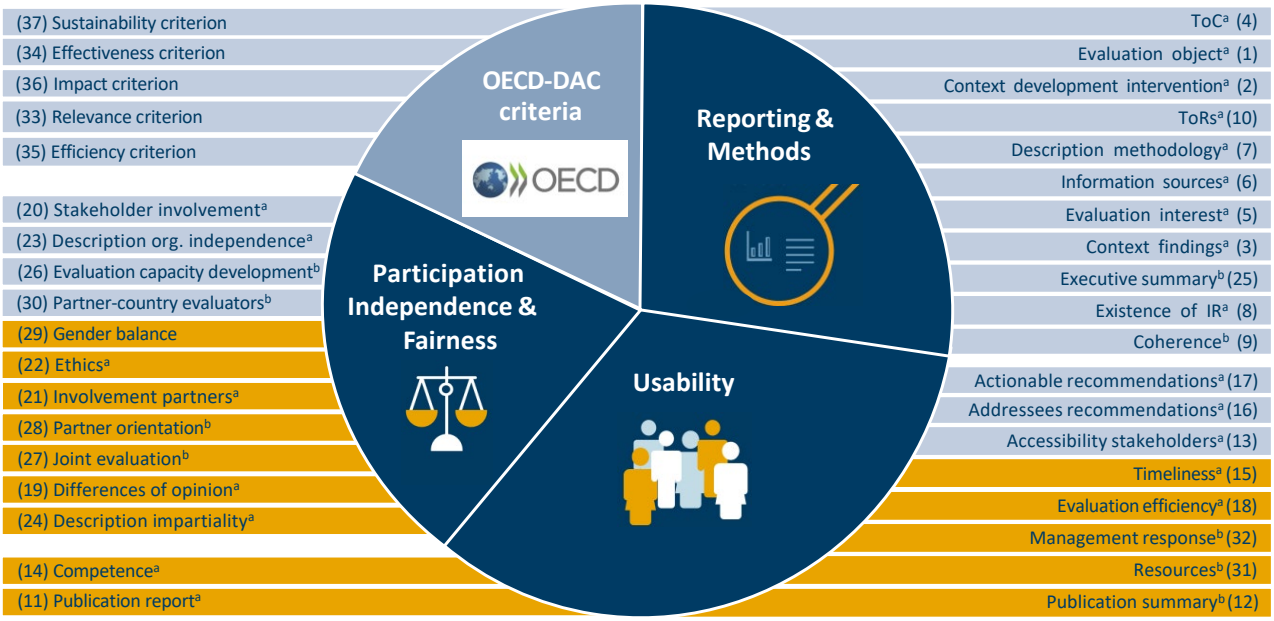
Reporting and methods. This standard cluster mainly comprises quality criteria that relate to either the presentation of information on the evaluation methodology (e.g. the "description of the Theory of Change" the "clarity of the information sources" or the "description of the methodological adequacy"), or the existence or content of selected evaluation documents (e.g. the "information content of the terms of reference", the "quality assurance with inception report" or the "information content of the executive summary"). Figure 3 shows the 32 quality criteria from the area of overlap between the OECD-DAC and DeGEval standards, and the area "OECD-DAC only", arranged according to their assignment to the three standard clusters.

Participation, independence and fairness. This standard cluster is primarily assigned quality criteria that concern the inclusion of different groups of people in the evaluation (e.g. the "involvement of internal and external stakeholders", the "inclusion of partner-country evaluators" and the "gender balance in the evaluation team"). Furthermore, this standard cluster covers aspects of independence (e.g. "description of the methodological adequacy") and ethical aspects (e.g. "evaluation ethics").

Usability. In this standard cluster, it is primarily (theoretical) usefulness that is relevant to the meta-evaluation, whereas (practical) use plays only a minor role, and benefit alone is not examined. Usefulness is defined as the evaluation's potential for use (and when used, the potential for benefit to emerge), which can be influenced before, during and after the evaluation's implementation.[29] Use is defined as the direct response to the evaluation content, which may be a written or spoken response to various aspects of the evaluation. Benefit describes the actual advantage that results from an evaluation (e.g. the increase in efficiency based on an evaluation's recommendations). In this standard cluster, quality criteria include e.g. the "publication of the evaluation report", the "timeliness of the findings" and the "actionable recommendations".

29 Thus, theoretically, virtually all quality criteria from all three standard clusters can be useful. The standard cluster "usability" contains those quality criteria that are directly and always linked to the usefulness of an evaluation.

Figure 3 Assignment of the 37 quality criteria to the standard clusters and OECD-DAC criteria



Source: DEval, authors' own graphic

Note: blue bar = quality criterion examined by evaluation; yellow bar = quality criterion examined at the level of the organisation across all evaluations; org. = organisational. [a] Quality criterion comes from the area of overlap between OECD-DAC and DeGEval standards; [b] Quality criterion comes from the "OECD-DAC only" area. The organisation-specific quality criteria and the quality criteria of the sustainability meta-evaluation are not shown, as they were not examined for all organisations.

2.2 Factors affecting application of the quality criteria

This section provides an overview of which factors (identified from literature and focus group discussions) are linked to or can explain the application of quality criteria. The identified factors are the basis for the regression analyses estimated in section 4.3 (evaluation question 3). Details on the definitions, the causal relationships, the hypotheses and the operationalisation of the factors can be found in section 2.2 of the online annex.[30]

The factors described below were analysed if they possessed three attributes, namely: 1) They had a clear cross-organisational definition. 2) It was possible to describe clear links to selected quality criteria. 3) Data were available either from the organisations or in secondary databases. To identify the factors, three focus group discussions[31] were conducted with the responsible officers of the involved organisations, and scientific and empirical literature was reviewed. The factors were then systematically categorised as either 1) country-specific, 2) evaluation-specific[32] or 3) organisation-specific (Table 3).

30 In the following sections, links between the factors and the quality criteria are formulated partially in causal terms based on the available information. The regression analyses estimated in section 4.3, however, are interpreted exclusively in terms of correlations. This is because the selected design did not allow for the measurement of rigorous causal relationships.

31 Focus group discussions are considered to be a proven instrument for addressing experiences and assessments on a specific topic, with organisation groups that are often heterogeneously structured, in a relatively short period of time (Morgan, 1999). Four further organisation-specific factors were also identified during a discussion at the reference group meeting on "initial findings and conclusions"

32 Some factors can be assigned to both the evaluation-specific and the organisation-specific categories. One example is the factor "quality assurance with inception report". However, since these factors are not always applied or implemented in the same way at the level of evaluations, they were categorised as "evaluation-specific" rather than "organisation-specific".

The country-specific dimension[33]

Empirical research and theoretical explanations indicate that "conflict/fragility" has no clear link to the application of the standard cluster "reporting and methods"; "pandemics" and the "cultural context" might influence all three standard clusters. There is evidence in the experience-based literature that both "conflict" and "pandemics" make data collection in evaluations and collaboration with local organisations more difficult due to access restrictions, an increased security risk and/or a lack of infrastructure (Church and Shouldice, 2002; OECD, 2012). However, an evaluation synthesis by DEval on the work of development cooperation in fragile contexts shows that "fragility" is not statistically correlated with methodological quality in evaluation reports (Wencker and Verspohl, 2019). So far, there has been no investigation of whether "fragility/conflict" is linked to quality criteria from the standard clusters "participation, independence and fairness" or "usability", or individual quality criteria from the standard cluster "reporting and methods", which are examined in this meta-evaluation. Another factor identified in the focus group discussions is the "cultural context". This could affect both the "reporting and methods" and the "participation, independence and fairness" of the evaluation if it played a major role in its design phase.

The evaluation-specific dimension

There is evidence in the experience-based literature that remote data collection has a positive impact on application of the quality criteria "inclusion of partner-country evaluators" and "incorporation of evaluation capacity development", and a negative impact on the "description of the methodological adequacy". Since the beginning of the COVID-19 pandemic, the conduct of remote evaluations has affected two things (among others): 1) cooperation in (international) teams/cooperation with partner-country experts, and 2) data collection methods. For example, cooperation with partner-country evaluators is of great importance when conducting remote evaluations, as they can identify the right stakeholders, are well networked and know the project or country context well (von Gumppenberg et al., 2022; Mäder, 2020; World Bank, 2020a, 2020b). Their involvement can also support partner-country "evaluation capacity development" (von Gumppenberg et al., 2022). Furthermore, remote or semi-remote evaluations often involve travel restrictions, meaning that data for an evaluation can no longer be collected on the ground. This may imply limitations in terms of "description of the methodological adequacy" (World Bank, 2020a; Hundt and Bräuer, 2021; Lange et al., 2020).

In meta-evaluations and in feedback from the focus group discussions, it emerged that the "information content of the terms of reference", "quality assurance with inception report" and "stakeholder involvement" are positively correlated with application of the quality criteria from the standard clusters "reporting and methods" and "usability". Several meta-evaluations show the positive links described above. One example is the link between "information content of the terms of reference", and a stronger engagement with the "evaluation object", the "partners" and "reporting and methods" (Caspari, 2010; FES, 2015; Queiroz de Souza, 2017; Silvestrini and Bäthge, 2019; Väth et al., 2022).

In other meta-evaluations and as reflected in the focus group discussions, the "number of evaluators", the "evaluator days"[34], the "competence of the evaluators", the "evaluation costs" and the "year of evaluation" affect application of the quality criteria in the standard cluster "reporting and methods", as well as other aspects of quality. Previous meta-evaluations prove empirically that a higher "number of evaluators" increases "methodological quality" (Freimann et al., 2016; Krämer and Almqvist, 2019), but not "usefulness" (Freimann und Krämer, 2017). "Reporting and methods" also increases with the "evaluator days". However, this only applies up to a certain number of "evaluator days", after which the correlation is no longer statistically significant or even declines (Freimann et al., 2016; Krämer and Almqvist, 2019). In existing meta-evaluations (Koy et al., 2016; Queiroz de Souza, 2017) and among the organisations involved,

33 The factors in this dimension are difficult for organisations to influence, but they may still explain why quality criteria are applied either more or less.

34 It should be remembered that the factors "number of evaluators" and " evaluator days" are interrelated or may be interdependent.

there is also discussion of whether the "evaluation costs" can theoretically influence the "methodological quality", but this has not yet been empirically investigated. Above all, this is based on the assumption that with more financial resources – i.e. higher costs –more means and thus for instance more "evaluator days" are available. Hageboeck et al. (2013) also state that the evaluation costs should be examined as a factor, but that this could not be implemented empirically due to low data availability. Hence there is currently no clear empirical evidence regarding the relationship between the "evaluation costs" and the application of quality criteria. Furthermore, it was noted in the focus group discussions that, based on experience, the "year of evaluation" is positively linked to the application of quality criteria, especially "reporting and methods", as the evaluation systems of the organisations usually develop or improve over time.

Box 1 Links between the quality criteria

In their two respective standards documents, the OECD-DAC and DeGEval place the individual quality criteria alongside each other on an equal footing. Nevertheless, it is likely that some quality criteria influence (or precede) others. Thus, in the literature as well as in the focus group discussions, some specific quality criteria are discussed as explanatory factors for the application of other quality criteria (e.g. it is argued that the "competence of the evaluators" affects the "description of the methodological adequacy") (Silvestrini and Bäthge, 2019). This can be assumed because in the OECD-DAC standards document the quality criteria are assigned to evaluation process phases and thus at least partly follow each other in time. In order to take this understanding and the previous findings into account, selected quality criteria were analysed both as factors and as affected quality criteria.

The organisation-specific dimension[35]

In dialogue with the involved organisations, the "size of evaluation units/desks", the "size of evaluation units/desks relative to number of evaluations", the "size of evaluation units/desks relative to size of organisation" and "evaluation activity" were identified as potential factors linked to the application of the quality criteria for the standard cluster "reporting and methods". From the point of view of the responsible officers of the involved organisations' evaluation units/desks, factors connected with the resources available to those responsible for evaluations – specifically the factors "size of evaluation units/desks", "size of evaluation units/desks relative to number of evaluations" and "size of evaluation units/desks relative to size of organisation" – are positively linked to the application of the quality criteria for the standard cluster "reporting and methods". Furthermore, a higher level of "evaluation activity" correlates positively with application of the quality criteria for the standard cluster "reporting and methods".[36] Table 3 provides an overview of the factors examined in the meta-evaluation.

35 Within this dimension, there are overlaps with the system assessment and its monitoring (Lücking et al., 2015), as aspects of the organisational context were also examined in the latter. However, the meta-evaluation examines the relationship between organisation-specific factors and the application of quality standards in evaluations, and not the extent to which organisation-specific aspects differ between organisations.

36 In the focus group discussions, "structured planning process", "evaluation unit in place" and "evaluation and learning culture" were also mentioned as organisation-specific factors. These could not be investigated as they did not display the three required attributes (cross-organisational definition, causal relationships and data availability). However, proxies were identified for some factors that did not fulfil one or more of these requirements. A proxy is a variable that is as similar as possible to the relevant variable in terms of content (e.g. for "competence evaluators", it was possible to use the "number of evaluators" [who have different expertise] and the average "daily rate of external evaluators" as proxies).

Table 3 Overview of the factors examined

Dimension	Factor	A1 – Definition	A2 – Causal relationships	A3 – Data availability	Proxy for the factor*	Level of measurement
Country-specific	Fragility	yes	yes	no	conflict	binary
	Cultural context	no	no	no	Social Capital Index	metric
	Pandemic	yes	yes	no	year 2020	binary
Evaluation-specific	Competence evaluators	no	yes	no	number of internal and external evaluators	metric
					daily rate of external evaluators	metric
	Quality assurance processes	no	yes	no	content ToRs	ordinal
					quality assurance with inception report	binary
					stakeholder involvement	ordinal
	Remote data collection	yes	yes	yes	/	ordinal
	Evaluation costs	no	yes	no	evaluator days relative to costs of development intervention	metric
	Year of evaluation	yes	yes	yes	/	metric
	Data availability among stakeholders**	no	no	no	/	

Dimension	Factor	A1 – Definition	A2 – Causal relationships	A3 – Data availability	Proxy for the factor*	Level of measurement
Organisation-specific	Structured planning process	no	yes	no	organisation	nominal
	Evaluation and learning culture	no	yes	no		
	Evaluation unit in place	no	yes	no		
	Size of evaluation units/desks	yes	yes	yes	/	metric
	Size of evaluation units/desks relative to number of evaluations	yes	yes	yes	/	metric
	Size of evaluation units/desks relative to size of organisation	yes	yes	yes	/	metric
	Evaluation activity	yes	yes	yes	/	metric

Source: DEval, authors' own table

*Note: A = attribute; A1–A3 refer to the attributes described at the beginning of this section that factors must display for them to be empirically investigated: A1 = a clear cross-organisational definition of the factor; A2 = clear causal relationships between the factor and selected quality standards; A3 = availability of data on the factors at the organisations. * If the factor could not be examined directly, for example because no precise or cross-organisational definition could be found, a proxy was defined instead that overlaps extensively with the original factor. "/" means that no proxy needed to be defined, and it was possible to analyse the factor directly (i.e. all attributes were evident). ** The factor "data availability" is covered by the factors "remote data collection" and "conflict". ToRs = Terms of Reference.*

3. METHODOLOGY

This chapter describes both the data and the analyses carried out. More detailed information on both can be found in section 6 of the online annex. The chapter also describes how the application of the quality criteria was rated, and what strengths and challenges existed in the meta-evaluation's methodology.

3.1 Data and data analysis

Selecting the involved organisations and evaluations

A two-step procedure was chosen for identifying the evaluations. First the organisations were identified, and then the evaluations of the selected organisations.

In the first step, the four official implementing (governmental) organisations – BGR, GIZ, KfW and PTB – and seven non-governmental organisations – CARE, DRK, DVV, EWDE, hbs, KAS and MISEREOR – were included in the meta-evaluation. The organisations were selected using four criteria. This was in order to cover the greatest possible structural heterogeneity of organisations (criteria 1 to 3), and enable the analysis of a sufficient number of evaluations per organisation (criterion 4). Criterion 1 related to the amount of BMZ funding (two organisations each with the highest and lowest average absolute BMZ funding per year). Criterion 2 related to the relative evaluation activity (one organisation each with the lowest and highest average BMZ funding per evaluation per year). Criterion 3 related to the budget item of the organisations (at least one organisation per budget item[37]). Selecting organisations (cases) by criteria that were designed to maximise heterogeneity (the diverse case method) did not allow any conclusions about the distribution of the application of quality standards across all non-governmental organisations in German development cooperation. However, it did allow conclusions concerning the application of quality standards for a wide range of different organisations. Organisations that were not included in the analysis could thus use the findings as a starting point for their own evaluation practice. The diverse case method thus comes closer to a representative study than other case selections with small samples (e.g. a selection based on homogeneous cases or extreme cases); (Seawright and Gerring, 2008).[38] Criterion 4 addressed the evaluation frequency of all organisations, in order to ensure a sufficient number of evaluations per organisation (approximately two evaluations per year).[39] Further details can be found in section 3.1 of the online annex.

In the second step, a random sample of 296 out of 576 evaluations was drawn, stratified by organisation and year. In total, the evaluation units/desks of the organisations were (co-)responsible for 849 evaluations in Germany during the period under review from October 2016 to December 2020[40]. The population encompassed 576 evaluations that the BMZ had either (co-)funded or in which a development intervention (co-)funded by the BMZ was examined (average coverage = 62 per cent). The sample drawn comprised 296 evaluations (on average 75.4 per cent of the population; Table 4). Due to the selected statistical

37 Organisations were included that are located in the "Civil society, municipal and private-sector engagement" section of Departmental Budget 23. Specifically, they fall under the BMZ budget items "Promotion of social structure projects that are important for development", "Promotion of projects of political foundations that are important for development", "Promotion of projects of churches that are important for development" and "Promotion of projects of private German organisations that are important for development" (BMF, 2020). Further information can be found in section 3.1 of the online annex.

38 Since appropriate criteria for a representative selection of non-governmental organisations were not available for the meta-evaluation, or could not have been determined within the time frame, the diverse case method was used.

39 For criteria 1 and 2, an organisation could only be selected once (for example, if an organisation received both the lowest BMZ funding and had the highest evaluation activity, it was only selected via one criterion; for the second criterion, the organisation in the next place was then selected). Two organisations were selected for criterion 1 and one organisation for criterion 2, as the amount of BMZ funding and the resources thus available for each evaluation were weighted higher than the organisation's evaluation activities. Criterion 3 was also taken into account, and criterion 4 was applied for all identified organisations.

40 Evaluations of hbs are available from January 2016 to October 2020, and of CARE and GIZ from January 2018 to December 2020.

parameters[41], the differences in the number of evaluations per organisation meant that organisations with a lower number of evaluations were more strongly represented in the sample studied. To be able to calculate the mean values for application of the quality criteria, for each organisation the ratio of evaluations from the sample compared to the population was taken into account by weighting (disproportionate sampling).

41 For each organisation, the number of evaluations in the sample was chosen in such a way that a margin of error of 10 per cent was not exceeded at a confidence interval of 95 per cent. The margin of error allows for a possible deviation of the findings of approximately 10 per cent from the real value. Since ordinal variables are used quasimetrically, a margin of error can be defined and a distribution assumption made. The confidence level indicates how certain the findings are. Together with the margin of error, it is thus possible, for example, to state that for an identified value of 40 per cent in the sample for a quality criterion, there is a 95 per cent probability that the value in the population is between 30 and 50 per cent. For an organisation with less than ten evaluations, all evaluations were examined.

Table 4 Involved organisations and number of evaluations

No.	Organisation	Budget item	TN of (co)managed evaluations[a]	POP[b]	SP[c]	Coverage[d]	POP per organisation as % of POP for all organisations	SP per organisation as % of SP for all organisations	SP as % of POP per organisation
1	BGR	FC/TC	31	21	18	67.7 %	3.6 %	6.1 %	85.7 %
2	CARE[e, f]	PGO	6	6	6	100.0 %	1.0 %	2.0 %	100.0 %
3	DRK	SST	64	20	17	31.3 %	3.5 %	5.7 %	85.0 %
4	DVV	SST	56	20	17	35.7 %	3.5 %	5.7 %	85.0 %
5	EWDE	Church	63	14	13	22.2 %	2.4 %	4.4 %	92.9 %
6	GIZ[e]	FC/TC	109	62	38	56.9 %	10.8 %	12.8 %	61.3 %
7	hbs[g]	PF	27	22	19	81.5 %	3.8 %	6.4 %	86.4 %
8	KfW	FC/TC	239	230	68	96.2 %	39.9 %	23.0 %	29.6 %
9	KAS	PF	39	20	17	51.3 %	3.5 %	5.7 %	85.0 %
10	MISEREOR	Church	158	123	55	77.8 %	21.4 %	18.6 %	44.7 %
11	PTB	FC/TC	57	38	28	66.7 %	6.6 %	9.6 %	73.7 %
		Total	**849**	**576**	**296**	**average: 62.5 %[h]**	**100.0 %**	**100.0 %**	**average: 75.4 %[i]**

Source: DEval, authors' own table

Note: eval. = evaluations; FC/TC = bilateral official Financial and Technical Cooperation; POP = population; TN = total number; PF = political foundation; PGO = private German organisation; SP = sample; SST = agency engaged in social improvement; [a] total number of all evaluations for which the (central) evaluation units/desks were (co-)responsible for accepting the report; [b] number of all evaluations which the (central) evaluation units/desks were (co-)responsible for accepting the report and which had been (co-)funded by BMZ in some form; [c] number of evaluations examined; [d] percentage of the population accounted for by the total number of evaluations for which the organisations are (co-)responsible; [e] figures refer to the years 2018 to 2020; [f] complete population studied, as fewer than ten evaluations were available; [g] figures refer to the period from January 2016 to October 2020; [h] percentage of the total number of evaluations of all organisations for which the latter are (co-)responsible accounted for by the population of all organisations is 67.8 per cent; [i] percentage of the population accounted for by the sample across the evaluations of all organisations is 51.4 per cent.

Data collection and analysis

The section below describes the collection and analysis of data for evaluation questions 1 (understanding of quality), 2 (application of quality standards) and 3 (links between selected factors and quality standards).

The understanding of quality

To determine the understanding of quality among the organisations involved, and which quality standards they were required to apply, the meta-evaluation examined documents of the organisations as well as relevant agreements with the BMZ. It also conducted interviews with the officers responsible for evaluations. It analysed the data obtained from these sources using a qualitative content analysis.

Application of quality standards

To examine the application of the quality criteria in the 296 evaluations, evaluation documents and other documents of the involved organisations were used. Furthermore, the responsible officers of the evaluation units/desks were surveyed online, in order to rule out a false negative assessment of application. In total, to answer evaluation question 2a (application of the OECD-DAC and DeGEval standards), approximately 1,000 evaluation documents (evaluation reports and annexes, ToRs, inception reports) and other documents at the organisational level (such as evaluation plans, guides and manuals for conducting evaluations, standardised templates for evaluation reports) were examined. In the inter-coding phase[42], for 14 of the 37 OECD-DAC and DeGEval quality criteria either very little information or none at all could be coded in the evaluation documents provided by the organisations. To avoid drawing erroneous conclusions about non-application, in a further step the responsible officers of the evaluation units/desks were therefore surveyed concerning the application of these quality criteria in their organisation. This was done online – thus flexibly in terms of time – and on a standardised basis.[43] One drawback of this procedure was that the online survey was conducted at the level of the organisation, and hence the findings for these quality criteria were not recorded at the level of the individual evaluation. One point of criticism that must be noted is that the responses were self-reported by the responsible officers of the evaluation units/desks, and could not be traced back to the individual evaluations.[44] The findings of the document analysis and the online survey are therefore shown in the graphics side by side – using different colours – in order to highlight the difference between the data sources.

42 The inter-coding phase was the first phase of coding, in which each quality criterion was coded by all coders in 10 per cent of the evaluations (N = 30 evaluations) (Döring and Bortz, 2016). In this phase, the application of a quality criterion was coded as "-99" if no information was identified in the evaluation. Following this phase, checks were performed to determine whether the information consistently could not be coded in more than 24 evaluations and at least nine organisations. If this was the case, the quality criteria were integrated into the online survey; if not, the -99 was transformed into a 1, thus classifying the absence of information in the evaluation documents as "barely applied". This transformation resulted in some quality criteria receiving rather low ratings (this applies in particular to the quality criteria "accessibility for stakeholders", "involvement of internal and external stakeholders", "description of the organisational independence of the evaluators" and "incorporation of evaluation capacity development"). Further details on the online survey can be found in section 3.4 of the online annex.

43 In contrast to the document analysis, the online survey gave the organisations the option of not providing any information on the application of specific quality criteria, without this being counted as "barely applied". This enabled the responsible officers of the evaluation units/desks to avoid having to make any inappropriate assessment of the application of a quality criterion across all evaluations. In eight cases, individual responses to the online survey were subsequently changed for good reason (in six cases the ratings were increased and in two they were decreased).

44 Since the quality criteria of the online survey were applied on average about 6 per cent less than the quality criteria of the document analysis, there was no reason to assume that the organisations systematically rated themselves more highly than they would have been rated by the objective coding.

Possible reasons for the lack of information on the application of the 14 quality criteria at the level of the individual evaluation can be 1) the complexity of the quality standards, 2) the lack of documentation of an (explained) non-application, 3) the fact that application was recorded in writing in evaluation documents that were not examined, or 4) the fact that application was documented at the organisational level.

1) Some quality criteria reflect complex content (e.g. "evaluation ethics", "evaluation efficiency", "sufficient resources available"). This means that a meaningful operationalisation across all organisations was almost impossible for these quality criteria, as up to eleven different operationalisation options can exist (for meta-evaluations within one organisation, criteria can be operationalised on an organisation-specific basis). The responses of the responsible officers of the evaluation units/desks in the online survey reflected the application of different operationalisation options at these points.
2) Discussion of the (non-)application of quality criteria has not yet found its way into evaluation practice. Thus, at the level of the individual evaluation, there is no evidence of the application of some quality criteria (such as "transparency of differences of opinion", "consideration of joint evaluations" and "partner-country orientation").
3) The application of quality criteria may have been written down in documents that were not made available to the evaluation team for various reasons (such as "competence of the evaluators" in the evaluators' application documents); or they may be found in channels that were not explored for all evaluations (such as "publication of the report" and "publication of the executive summary" on the organisations' websites).
4) The organisations had documented the application of the quality criteria at the level of the organisation, but not at the level of the individual evaluation (for example, "existence of a management response", "evaluation efficiency").

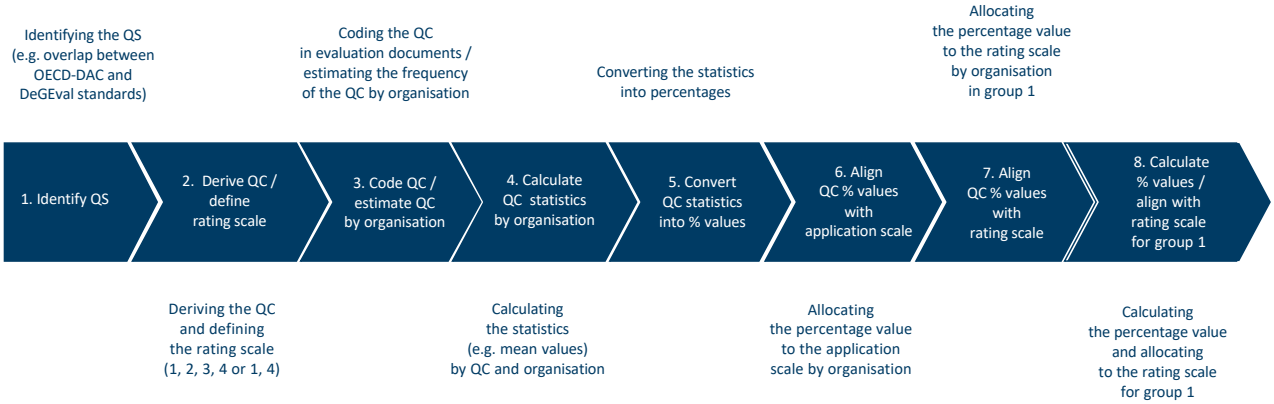
To analyse the application of the quality criteria derived from the OECD-DAC and/or the DeGEval standards, quantitative content analyses were performed and descriptive statistical methods were used. For the analysis, the quality criteria described in the analysis grid (section 2.1) were coded on the basis of ordinal scores (1 = "not applied", 2 = "barely applied", 3 = "largely applied", 4 = "applied") or binary scores (1 = "not applied", 4 = "applied"). One advantage of quantitative content analysis is that the application of quality criteria could be recorded and analysed both in a manner that is as intersubjectively evident as possible and at the level of the individual evaluation (Döring and Bortz, 2016).[45] Values were also assigned to the average frequencies of application of the quality criteria reported by the responsible officers of the evaluation units/desks in the online survey (1 = "never", 2 = "rarely", 3 = "partly", 4 = "largely/frequently", 5 = "always"). For each quality criterion, different statistics (e.g. mean values, medians, maximum and minimum values) were calculated for each organisation. These statistics were subsequently converted into percentages[46] and assigned to the predefined thresholds (section 3.2). Subsequently, in accordance with the respective requirements (section 3.2) a cross-organisational mean value was calculated from the individual mean values of the organisations for both groups, and rated for group 1 (Figure 4). Since the average cross-organisational mean value per quality criterion was calculated on the basis of the mean values per organisation, this value is independent of the number of evaluations of the individual organisation. To

[45] The average inter-coding agreement was good, with a value of 0.77. The lowest value was 0.63, the highest 0.97. The Krippendorff's alpha coefficient can range from 0.00 (lack of reliability) to 1.00 (perfect reliability). Values of $0.80 \leq \alpha \leq 1.00$ are considered good and values of $0.67 < \alpha \leq 0.80$ are considered acceptable; Krippendorff, 2012: 241). Detailed information on the procedure and on calculation of the inter-coding agreement can be found in section 3.3 of the online annex.

[46] For the analysis, the metrics were calculated taking into account the disproportionate sampling (using design weights) and converted to percentages using normalisation. Since rating scores are binary or ordinal, the percentages represent values that were not defined in the rating scores. In total, 19 quality standards comprised one quality criterion, five quality standards comprised two quality criteria, one quality standard comprised three quality criteria and one comprised five. For the seven quality standards that comprised several quality criteria, the value was calculated separately. Since it was difficult for the evaluation units/desks to estimate precisely the mean values across the evaluations, the quality criteria of the document analysis were included in the quality standard value at 100 per cent and those of the online survey at 50 per cent.

calculate the standard cluster "reporting and methods", a summated index (value across all quality criteria) and factor values were determined.[47]

Figure 4 Steps in the process of rating the application of quality criteria



Source: DEval, authors' own graphic

Note: QC = quality criterion; QS = quality standard. Steps 1 and 2 are described in section 2.1 (understanding of quality, analysis grid and standard clusters), steps 3 to 8 in the previous section.

The explanations given for the application/non-application of quality criteria, which were investigated at the level of the organisations, were examined using a qualitative content analysis. Their documentation was examined with a quantitative content analysis. The qualitative content analysis (Kuckartz, 2014), which structures content, enabled a systematic analysis of written information from the online survey. This delivered a comprehensive picture of formal and actual evaluation practice. Furthermore, a quantitative content analysis of the organisations' documents was used to examine whether and to what extent the application/non-application of these quality criteria was prescribed. In the present meta-evaluation, a non-application that was explained in an organisation's documents was rated as "achieved". This occurred for two quality criteria in five organisations (out of 14 examined quality criteria in the online survey for eleven organisations each, this corresponds to approximately 2.6 per cent of cases). The explained non-application at the level of the organisation was assessed positively in terms of application, as this was a transparent and logical way to manage the quality criteria that affected all evaluations of the organisation.[48]

For four organisations, a total of eleven organisation-specific quality criteria were coded in relation to binary or ordinal scores. To answer evaluation question 2b (application of organisation-specific quality criteria), organisation-specific quality criteria (up to three per organisation, described in section 2.1) were coded in the evaluation documents of the organisations and subsequently calculated and evaluated like the OECD-DAC and DeGEval quality criteria.[49]

47 The values of all quality criteria in the standard cluster "reporting and methods" were included in the summative index in equal parts – and equally weighted. For the standard clusters "participation, independence and fairness" and "usability", no summative indices were calculated due to the small amount of information on the quality criteria at the level of the individual evaluation ("participation, independence and fairness": four out of eleven quality criteria; "usability": three out of ten quality criteria). Furthermore, factor scores were determined for the standard cluster "reporting and methods". These represent a weighted summary of the quality criteria and were determined by means of an exploratory factor analysis. Exploratory factor analysis is a statistical method that was used to examine whether the quality criteria of the standard cluster "reporting and methods" can also be summarised empirically (Backhaus et al., 2015; Brown, 2006).

48 In this approach, the quality of the explanation given was not rated. Alternatively, (explained) cases of non-application might also be described separately from the application of the quality criteria.

49 The intra-coder agreement (reselection and reclassification of at least 10 per cent of the evaluations per organisation at two different points in time) averaged a good value of 0.91. The lowest value was 0.68, and the highest 1.00. Intra-coder agreement was calculated if the evaluations were predominantly coded by one person. Further details on the intra-coder agreement can be found in section 3.3 of the online annex.

The evaluations of GIZ and KfW were also examined in relation to the quality criteria of the sustainability meta-evaluation. To answer evaluation question 2c (application/repeated application of the quality criteria from the sustainability meta-evaluation), 15 quality criteria from the sustainability meta-evaluation of Noltze et al. (2018) were used. Eight quality criteria were already included as quality criteria in the OECD-DAC and DeGEval analysis grid and were merely transformed (for example, by converting four scores into two). The remaining seven were recoded in the current evaluations (1 = "not applied", 4 = "applied".[50] Subsequently, the difference in findings between the sustainability meta-evaluation (t0) and the present study (t1) was examined using a structural equation model (Weiber and Mühlhaus, 2010). This enabled the simultaneous calculation of the factor "time" (t0 versus t1; independent variable) along with several quality criteria (dependent variables). Furthermore, the statistical links between the quality criteria were taken into account.

Links between selected factors and the quality standards

To investigate links between selected factors and the application of the quality criteria, multivariate regression analyses were estimated. Regression analyses allow the identification of statistical links[51] between the factors (independent variables) and the quality criteria as well as the standard cluster "reporting and methods" (dependent variables; Backhaus et al., 2011). The factors examined were either 1) country-specific, 2) evaluation-specific or 3) organisation-specific (Table 5). The information for each factor was obtained from data submitted by the organisations or obtained from secondary databases.

3.2 Evaluating the application of the quality standards

For organisations required to apply the quality criteria (group 1), application of the quality criteria was both analysed and rated. For organisations with no such requirement (group 2), application of the quality criteria was only analysed. Both groups were analysed with regard to their degree of application of the quality criteria for the OECD-DAC and the DeGEval standards. Group 1 was also rated. For the OECD-DAC criteria (BMZ, 2006)[52], the organisation-specific criteria and the quality criteria of the sustainability meta-evaluation, only organisations of the group 1 were analysed and rated (Table 5).

50 To ensure the reliability of the quality criteria between the meta-evaluations, a three-stage procedure was selected: 1) The additional quality criteria were discussed with reference to selection and classification in a discursive procedure between a coder of the sustainability meta-evaluation and the coders of the present meta-evaluation. 2) A reduced sample of eleven evaluations (approximately 10 per cent of 106 evaluations) was randomly drawn and coded by the coders of this meta-evaluation. The inter-coder agreement was 1.00. 3) Since the 106 evaluations were almost exclusively coded by one person, the intra-coder agreement was also calculated. This also remained at a value of 1.00 throughout. Details can be found in section 3.3 of the online annex.

51 As no rigorous impact assessment was conducted, the findings of the multivariate regression analyses were interpreted correlatively. Individual organisations were included in the regression analyses to control for differences between them. The findings are valid across all organisations, regardless of the number of evaluations they contributed.

52 All organisations in the area "OECD-DAC criteria" were assigned to group 1, as they had committed to applying the OECD-DAC criteria in their organisational documents.

Table 5 Assignment of the organisations to groups 1 and 2 by area

Area	BGR	CARE	DRK	DVV	EWDE	GIZ	hbs	KAS	KfW	MISEREOR	PTB
Overlap between the OECD-DAC and DeGEval standards	G1	G2	G2	G1	G1	G1	G1	G1	G1	G1	G1
OECD-DAC only	G1	G2	G2	G2	G2	G1	G1	G1	G1	G2	G1
OECD-DAC criteria	G1	G1	G1	G1	G1	G1	G1	G1	G1	G1	G1
Organisation-specific QCs	/	/	G1	/	G1	G1	G1	/	/	/	/
QCs for the sustainability meta-evaluation	/	/	/	/	/	G1	/	/	G1	/	/

Source: DEval, authors' own table

Note: QC = quality criterion; G1 = organisations with a requirement (application of the quality criteria was analysed and rated); G2 = organisations with no requirement (application of the quality criteria was analysed, but not rated); / = no analysis

The thresholds for quantifying application of the quality criteria were defined in dialogue with the reference group. The threshold represented the ex ante assessment of when a quality criterion was considered to be barely applied, partly applied, largely applied or applied in an evaluation. When setting the thresholds in increments of 25 per cent (0 ≤ 25 per cent = "barely applied", > 25 ≤ 50 per cent = "partly applied", > 50 ≤ 75 per cent = "largely applied", > 75 ≤ 100 per cent = "applied"), the evaluation took into account the fact that the quality standards are to be understood as maximum standards.[53] It is therefore understandable "that not all standards can be realised in full" (DeGEval, 2016: 28). Since application of the quality criteria of the sustainability meta-evaluation was now being analysed once again, at this point it was postulated that the application of the quality criteria had improved since that meta-evaluation.

The rating was based on the defined thresholds for application, and by adding the extreme values 0 (not achieved) and 100 (exceeded) for group 1. To standardise ratings in DEval evaluations, the DEval rating scales (2020)[54] were drawn up. Based on this, six evaluation categories were defined based on the thresholds: 0 = "not achieved", > 0 ≤ 25 per cent = "barely achieved", > 25 ≤ 50 per cent = "partly achieved", > 50 ≤ 75 per cent = "largely achieved", > 75 < 100 per cent = "achieved", 100 per cent = "exceeded" (Figure 5 and section 7.3 in the annex to the report). In order to adequately rate the application of the quality criteria, knowledge of their non-application is also required.[55] Where documentation of non-application was lacking, this was rated as "not achieved". It should be noted that a lack of documentation need not mean that the quality criterion was in fact not applied in the evaluation. Nonetheless, since this was not verifiable, it was rated as "not achieved". This approach meant that the organisations received rather low ratings for the application

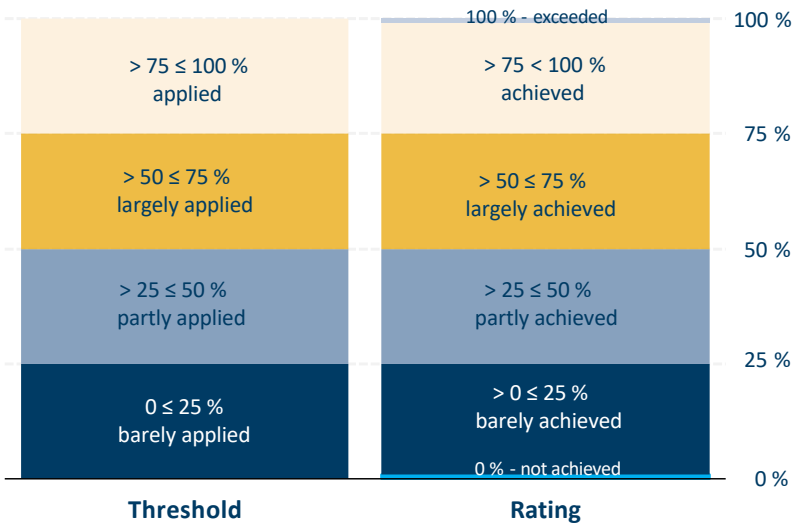
53 Since there were no scientific or empirical templates for spacing the benchmark threshold values in increments of 25 per cent, this segmentation is based on the logicality and transparency of the reasoning, the feasibility of determining the values and the acceptance by the organisations involved, VENRO and the BMZ.

54 DEval uses largely standardised rating scales to facilitate ratings for organisations involved in evaluations.

55 This is stated explicitly only in the DeGEval standards, but can also be transferred to the OECD-DAC standards, since without documentation of non-application neither an adequate assessment of application nor cross-sectional analyses are possible. Accordingly, the DeGEval standards (2016: 29) state: "The partial or complete non-application of individual standards should always be documented and explained clearly and explicitly, for example in reporting".

of certain quality criteria. The present meta-evaluation makes this transparent in the description of the findings for the relevant quality criteria.

Figure 5 The relationship between threshold and rating



Source: DEval, authors' own graphic

3.3 Strengths and challenges in the methodology

The selection of the organisations based on their structural heterogeneity enabled the meta-evaluation to analyse and describe the application of individual quality criteria across a correspondingly wide range. Accordingly, non-governmental organisations that were not involved can locate themselves within this range and use the findings of the meta-evaluation for themselves. The findings of this meta-evaluation are valid for the involved organisations. The official implementing organisations are fully described. The selection of involved non-governmental organisations is not representative. However, this does not mean that individual findings might not also apply to other non-governmental organisations and therefore be useful.[56] Since the organisations were selected on the basis of their structural heterogeneity, the distribution of application for all organisations of non-governmental development cooperation is not known. Nonetheless, the range of possible application or engagement with the quality standards is described.

The transferability of the findings to the population of evaluations by organisation is ensured by the selected statistical parameters for sampling. The findings are not transferable to other evaluation types of an organisation. There is no systematic analysis of whether or to what extent the findings on the application of quality criteria apply to other evaluation types that were not examined. It is possible that evaluations implemented using the same processes in an organisation, but funded from other sources, for instance, may show similar findings. On the other hand, it is to be assumed that evaluations conducted using other processes (e.g. decentralised evaluations) are more likely to produce different findings (BMZ, 2021; Koy et al., 2016).

Since the analysis grid was derived from the OECD-DAC and DeGEval standards, it can also be used by other organisations. The fact that the analysis grid is based on the OECD-DAC and DeGEval evaluation standards documents (which are also used in the BMZ Evaluation Policy BMZ, 2021), makes the meta-evaluation more useful. The analysis grid of this meta-evaluation can thus be used in the future as a basis for preparing an analysis grid based on the BMZ Evaluation Policy.

56 When the meta-evaluation began, no attributes that would have enabled a representative sampling of all non-governmental organisations were known. Nor could they be identified within the time frame of the meta-evaluation.

When analysing the quality criteria from the online survey, there were limitations with regard to the triangulation of methods. In light of the cost-benefit ratio, however, adding a further data collection method would not have been appropriate. For those quality criteria that were collected through the survey of responsible officers of the evaluation units/desks (self-reported data across all evaluations at the level of the organisation), the assessments of the former evaluators involved in the evaluation could have been used to triangulate the data. However, this meta-evaluation covered a large number of evaluations from different organisations over a period of more than four years. It was therefore beyond the scope of this study to interview former evaluators from those evaluations instead of or in addition to the officers responsible for the evaluations. Particularly due to staff turnover, it would not have been possible to locate and interview some of the evaluators. The sample size would thus have been reduced. Furthermore, a survey would also have entailed risks, and thus would not have been sufficiently proportionate to the benefits. For the sake of transparency, the findings from the document analysis and the online survey are presented separately from each other in this report (see section 4.2.1).

Generally speaking, there are limits to measuring some quality criteria. For certain quality criteria, it would take a lot of effort to investigate a "good" application. There are quality criteria that can only be examined in depth with a great deal of effort. For example, for the quality criteria "stakeholder involvement" and "accessibility for stakeholders", both the appropriate number of stakeholders who can be involved and the intensity of their involvement in the various evaluation phases are difficult to determine. This also applied to the quality criterion "description of the methodological adequacy". For example, this was operationalised as follows: "The quality criterion is achieved if a) the appropriateness of the methods used is plausibly explained, and b) limitations of the methodological approach are discussed." Consequently, when coding no check was made of whether the methods described were actually appropriate; it was a question of whether the explanations given were plausible and whether the limitations were discussed. Thus, the selected procedures for operationalising the quality criteria do capture relevant aspects of the criteria, but do not necessarily reflect the depth of the underlying quality requirement. The quality criteria "description of the impartiality of the evaluators" and "description of the organisational independence of the evaluators" would also require a comprehensive qualitative analysis of the contracting and recruitment processes of the organisations, and of the convictions of the evaluators, in order to be able to assess impartiality and independence in accordance with the understanding of quality. One possible way of improving assessments in the future would be to record the necessary information through feedback from the evaluators or responsible officers of the evaluation units/desks at the time of the evaluation. The operationalisation of these quality criteria could then be defined in more precise detail accordingly.

For the cross-organisational meta-evaluation, the way in which the quality criteria were operationalised was developed across all organisations. Some of these operationalisation choices did not match the evaluation practice of all organisations. This meant that for these organisations, application received a lower score than it would have if the criteria had been operationalised otherwise. There are quality criteria whose operationalisation could be derived relatively clearly from the standards documents (for example, "information content of the terms of reference" as regards the inclusion of various aspects). For other quality criteria, however, there was more leeway (for example, "quality assurance processes"). Since it was in the interest of the present meta-evaluation to generate cross-organisational findings, operationalisation was defined in most cases. Accordingly, there is a conflict of objectives here between the meta-evaluation's interest in the application of selected quality criteria across organisations, and the heterogeneity of the application of the quality criteria. In the findings (section 4.2.1), the fact that certain quality criteria can be applied in several ways is made transparent.

Due to an inconsistent understanding of the measurement of "evaluation costs" across organisations, it was only possible to analyse explanations of the application/non application of quality standards to a limited extent. It was not possible to determine the costs of evaluations or to examine this factor in the analysis. Accordingly, a proxy had to be used to analyse the links between the evaluation costs and the application of selected quality standards. Without a common understanding or clear definition of evaluation costs, no insights can be gained in this regard.

Examining the quality criteria of the sustainability meta-evaluation once again enabled the present meta-evaluation to look at the difference in the application of the quality criteria by the GIZ and KfW over time. It thus provided evidence of the extent to which efforts and measures within organisations were able to improve the application of quality criteria. The difference in findings over time provides evidence of the extent to which application of the quality criteria of the sustainability meta-evaluation was improved, for example, by means of internal organisational reforms and the support of external actors (BMZ and DEval). It also brings to light challenges associated with a longitudinal study (such as raising the thresholds, and where necessary making appropriate adjustments to quality criteria over time). As the repeated implementation of meta-evaluations has not yet found its way into regular evaluation practice in most organisations, these findings also point to learning opportunities for other organisations.

4. FINDINGS

The findings chapter is divided into three sections. In each section, one of the three evaluation questions is addressed. First of all, the understanding of the quality of evaluations that exists among the involved organisations is described (evaluation question 1). This is followed by a description of the extent to which the involved organisations applied the OECD-DAC and DeGEval quality criteria (2a), the organisation-specific quality criteria (2b) and quality criteria of the sustainability meta-evaluation (2c). Finally, the links between the identified explanatory factors and the application of the quality criteria are described (3). In the annex to the report, the findings for the four official implementing organisations BGR, GIZ, KfW and PTB are also presented and classified at the level of the individual organisation (section 7.1).

4.1 The understanding of quality among the organisations involved

The first section describes the involved organisations' understanding of quality during the period of the analysis. Box 2 shows the general conclusion and the main findings.

Box 2 General conclusion on the understanding of quality

What understanding of evaluation quality do the involved German development cooperation organisations have? (Evaluation question 1)

The involved organisations' understanding of quality was predominantly based on the OECD-DAC and/or DeGEval standards and, where available, organisation-specific quality standards. When the meta-evaluation began, the involved organisations had in some cases not addressed these quality standards systematically.

- Six of the eleven organisations were required to apply both the OECD-DAC and the DeGEval standards (BGR, GIZ, hbs, KAS, KfW and PTB), three were required to apply only the DeGEval standards (DVV, EWDE and MISEREOR) and two organisations had no requirement (CARE and DRK). All eleven organisations were required to apply the OECD-DAC criteria, and four organisations also included organisation-specific quality standards (DRK, EWDE, GIZ and hbs). (Finding 1)
- At the beginning of the meta-evaluation, the OECD-DAC and/or DeGEval quality standards, and where available organisation-specific quality standards, were in some cases not fully known within the involved organisations, nor were they systematically prescribed in organisational documents and operationalised. (Finding 2)
- The BMZ requirements concerning the application of the quality standards varied in the relevant budget items during the period under review – in some cases the OECD-DAC standards were marked as mandatory, in others no requirements were specified.[57] (Finding 3)

The involved organisations were subject to different requirements concerning the application of the OECD-DAC, DeGEval and/or organisation-specific quality standards. For the most part, these requirements first had to be clarified with the organisations at the beginning of the meta-evaluation. With the exception of CARE and DRK, all organisations were required to apply the DeGEval standards by virtue of membership[58] and/or the description in their organisational documents; for the GIZ and KAS[59] this was in addition to the OECD-DAC standards. The four implementing organisations and two political foundations were also required to apply the OECD-DAC standards by the "Guidelines for bilateral Financial and Technical Cooperation with

57 The BMZ Evaluation Policy published in 2021 states in particular that both the OECD-DAC and the DeGEval standards are binding for the official implementing organisations, and provide guidance for the non-governmental organisations.

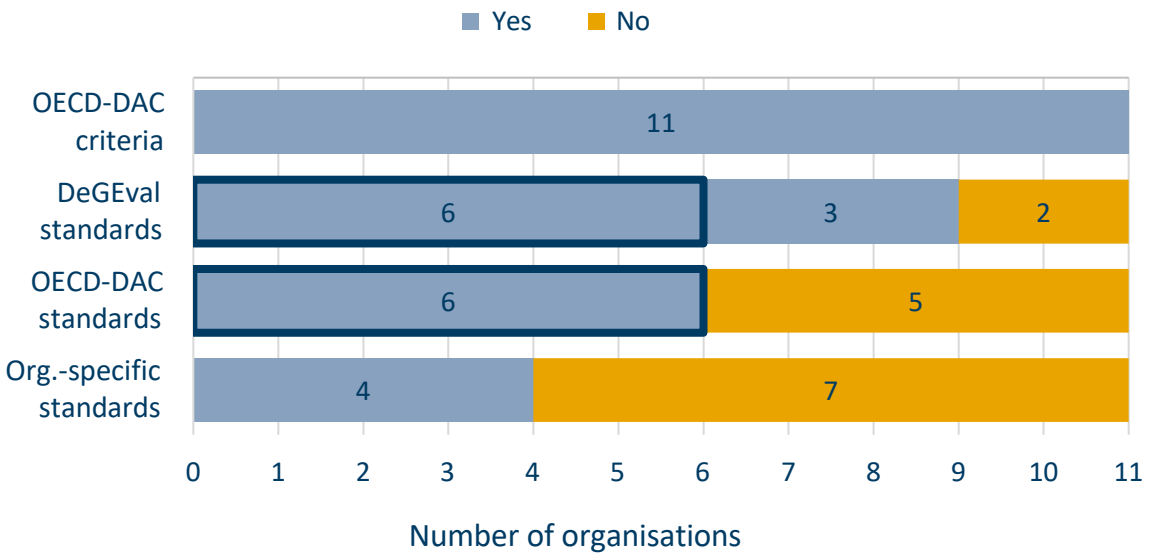
58 Since 2009, an organisation has committed itself to the DeGEval standards by signing the declaration of accession. Organisations that have signed the (current) declaration of accession since its adoption on 21 September 2016 have committed themselves to the current revised version of the DeGEval standards (DeGEval, 2021). For involved organisations that joined DeGEval before 2016 or 2009, a possible commitment to the DeGEval standards was ascertained by consulting the organisational documents, and/or in dialogue with the responsible officers of the evaluation units/desks.

59 The KAS committed to the OECD-DAC standards from 2016 to 2019.

cooperation partners of German development cooperation" (BMZ, 2021b)/the "Funding guidelines for political foundations" (BMZ, 2016). The BMZ had not issued any instructions for the organisations covered by the budget items "private German organisations", "agencies engaged in social improvement" and "Churches". In total, during the period under review six organisations were required to apply both standards documents. In the BMZ instructions, the focus was on the OECD-DAC standards. The organisational documents, on the other hand, more frequently described a requirement to apply the DeGEval standards. All eleven involved organisations were (and are) required by their organisational documents to apply the OECD-DAC criteria integrated in the OECD-DAC standards (Figure 6).[60] The BMZ Evaluation Policy (2021a), which was not yet in force at the beginning of this meta-evaluation, has further formalised and specified the BMZ's understanding of quality and its requirements concerning the application of quality standards. Here the BMZ refers in particular to application of the OECD-DAC standards and the DeGEval standards.

The eleven identified organisation-specific quality criteria of the DRK, EWDE, GIZ and hbs focused on content-related, methodological and partnership-related aspects of evaluations. Organisation-specific quality standards complemented the OECD-DAC and DeGEval standards, and were of particular importance to the organisations. Two other organisations referred to organisation-specific quality standards that overlapped with quality standards of the existing analysis grid. These thus represented a focus on selected internationally recognised quality standards rather than additional quality standards. The organisation-specific quality standards that were not covered by the standards documents included, for example, "partner inclusion in the recommendations" or the "contribution analysis". When reviewing the organisational documents, it turned out that organisation-specific quality standards – where they existed – were only partially prescribed explicitly. Hence their existence was subject to a certain scope for interpretation.[61]

Figure 6 Number of organisations required to apply selected quality standards



Source: DEval, authors' own graph

Note: The dark blue border indicates that six organisations (implementing organisations and political foundations) were required to apply both the OECD-DAC and the DeGEval standards. Application of the OECD-DAC criteria is a quality standard of the OECD-DAC standards.

[60] Whether or not an organisation was required to apply quality standards also determined whether that organisation was allocated to group 1 or 2 in the presentation of the findings (Table 5, section 3.2).

[61] For more detailed information on the organisation-specific quality criteria, see section 3.3 of the online annex.

4.2 Rating of the application of the quality criteria

The second section describes the application of the quality criteria in four subsections: 1) overlap of the OECD-DAC, the DeGEval and the OECD-DAC only standards, 2) the OECD-DAC criteria, 3) organisation-specific quality criteria and 4) quality criteria of the sustainability meta-evaluation (including difference in application over time). Box 3 provides an overall conclusion.

Box 3 Overall conclusion

A positive picture emerged in that the involved organisations on average largely applied the OECD-DAC, the DeGEval and organisation-specific quality standards and quality criteria of the sustainability meta-evaluation. This shows that application of the quality standards was actually part of the evaluation practice of the organisations. The findings for the organisations were very heterogeneous. Furthermore, it emerged that the quality standards had largely not yet been fully identified by the organisations in the organisational documents, and that their application/non-application had not been systematically prescribed. This was also the case for the traceability of the application/non-application of some selected quality standards at the level of the individual evaluation.

4.2.1 OECD-DAC and DeGEval quality criteria

This first sub-section shows the strengths and weaknesses in the application of the overlap between the OECD-DAC and DeGEval standards, and the OECD-DAC only standards. Box 4 shows the general conclusion and the main findings.

Box 4 General conclusion on application of the OECD-DAC and DeGEval quality standards

To what extent are strengths and weaknesses evident in the application of the OECD-DAC and the DeGEval standards in the evaluations of the involved German development cooperation organisations? (Evaluation question 2a)

Overall, a positive picture emerged regarding the application of the OECD-DAC and DeGEval standards. The involved German development cooperation organisations applied the quality standards in about two thirds of their evaluations. This was also the case – to a somewhat lesser degree – for organisations not required to apply the quality standards. The differences in the application of the quality standards between the organisations sometimes varied widely. This was to be expected, given the selection criteria for the inclusion of the involved organisations in the sample. It was thus possible to obtain a heterogeneous picture across the varying degrees of application.

It should be noted that, for various reasons, the application of some quality standards was recorded not at the evaluation level, but at the organisational level. This might be due to (i) the way in which the meta-evaluation chose to operationalise some quality criteria, (ii) lack of documentation on application/non-application, or (iii) the fact that application was documented exclusively at the organisational level rather than at the level of the individual evaluation. There is a clear need for improvement here, as without information, an external investigation of an (explained) application/non-application of the quality standards at the level of the evaluation was only possible to a limited extent. It was thus not possible to trace whether a quality standard was either not applied (with or without explanation), or was applied but was not documented.

a) Strengths were evident in the application of the quality standards.[62] (Finding 4)

- On average, 68 per cent of the 37 quality criteria were applied. About three quarters of the quality criteria (29 out of 37) were on average largely applied/applied and one quarter partly or barely applied. The three most frequently applied quality criteria were "description of the evaluation object (1)", "evaluation ethics (22)" and the average "application of the OECD-DAC criteria (33-37)". (Finding 4.1)
- There were sometimes major differences between organisations in their application of the quality criteria. (Finding 4.2)
- Even organisations with no requirement to apply the quality criteria did apply them in their evaluation practice, the average being 61 per cent. (Finding 4.3)

b) Weaknesses were found in the traceability of the application/non-application of some quality standards at evaluation level, and in the identification and systematic description of relevant quality standards in organisational documents. (Finding 5)

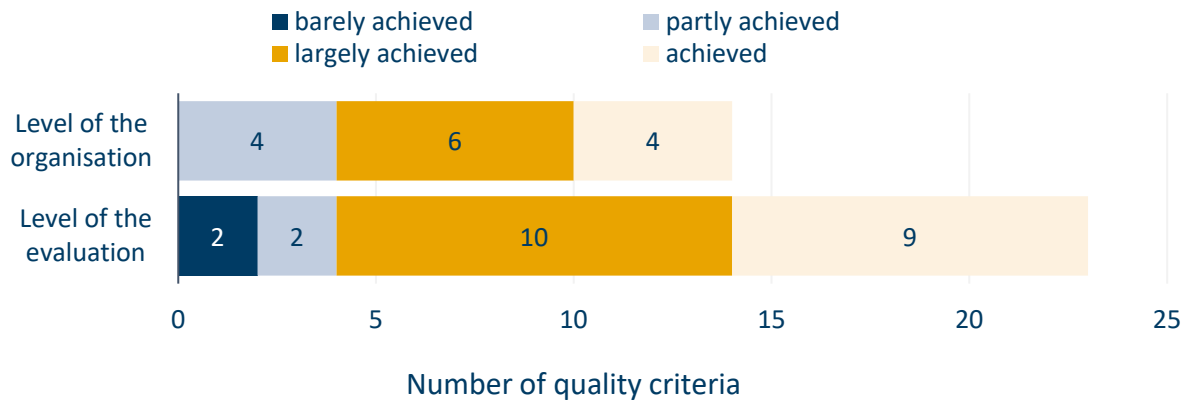
- The quality criteria that were required for the organisations were in some cases not identified, and sometimes only partially systematically prescribed in the organisational documents. (Finding 5.1)
- About a quarter of the quality criteria were partly or barely applied. The four least applied quality criteria were "inclusion of partner-country evaluators (30)", "incorporation of evaluation capacity development (26)", "description of the impartiality of the evaluators (24)" and "partner-country orientation (28)". (Finding 5.2)
- The application of approximately 38 per cent of the quality criteria (14 out of 37) was captured not at the level of the individual evaluation, but rather across all evaluations for an organisation by means of an online survey. One reason for this was that the non-application of the quality criteria was almost never recorded in the individual evaluations. (Finding 5.3)
- The non-application of selected quality criteria was explained by organisational or evaluation-specific factors, or by ways of applying them that did not match the ways in which application was operationalised in the present meta-evaluation. (Finding 5.4)

Overarching findings

On average, 68 per cent of the 37 quality criteria were applied, and thus rated as "largely achieved". There were clear differences in application between the organisations. Organisations not required to apply the quality criteria also applied them. About three quarters of the quality criteria were on average largely achieved or achieved (N = 29), one quarter partly or barely (N = 8; Figure 7). The three quality criteria achieved to the highest extent were the average "application of the OECD-DAC criteria (33-37)", "description of the evaluation object (1)" and "evaluation ethics (22)". The four achieved to the lowest extent were "inclusion of partner-country evaluators (30)", "incorporation of evaluation capacity development (26)", "description of the impartiality of the evaluators (24)" and "partner-country orientation (28)". Furthermore, for about 57 per cent of the quality criteria (N = 21), the minimum and maximum values for the organisations in group 1 (organisations required to apply quality criteria) differed by more than 50 per cent. The absolute values for group 1 were at least 20 per cent higher for seven quality criteria, and lower for four quality criteria, than the values for group 2 (organisations not required to apply quality criteria).

62 In this section, the findings are mostly presented at the level of the quality criteria. Exceptions to this are overarching findings, and the presentation of findings for quality standards comprising two or more quality criteria.

Figure 7 Number of OECD-DAC and DeGEval quality criteria by degree of achievement



Source: DEval, authors' own graph

Note: N = 37 quality criteria. Level of the organisation = data collected by online survey across all evaluations; level of the evaluation = data collected by document analysis for each individual evaluation. No quality criterion received an average rating of 0 per cent (not achieved) or 100 per cent (exceeded).

The average finding across the involved organisations showed that most of the quality criteria which were achieved or largely achieved were in the standard clusters "usability" and "reporting and methods". In the standard cluster "participation, independence and fairness", partner-related quality criteria received lower scores. In the standard cluster "participation, independence and fairness", the quality criteria "partner-country orientation (28)", "incorporation of evaluation capacity development (26)" and "inclusion of partner-country evaluators (30)" scored among the lowest. Thus overall, a rather low involvement of partners was evident. The quality criterion "involvement of internal and external stakeholders (20)" was on average "largely achieved" by group 1 and represented a positive aspect of participation. Organisations in group 2 showed higher scores on average for these quality criteria. The quality criterion "involvement of internal and external stakeholders (20)" was on average "largely achieved" by group 1, reflecting a positive aspect of participation.

Application of approximately 38 per cent of the quality criteria (14 out of 37) was recorded not at the level of the individual evaluation, but rather across all evaluations for an organisation through an online survey. The lack of recording at evaluation level may be due, among other things, to the fact that either 1) the application/non-application of these quality criteria was not transparently described by the organisations; 2) application was prescribed in organisational documents that were not made accessible to the evaluation team; 3) application was prescribed not at evaluation level, but exclusively in the organisational documents, or 4) there was no appropriate cross-organisational operationalisation. The application of all quality criteria in the standard cluster "reporting and methods" and in the area "OECD-DAC criteria" was recorded at the level of the individual evaluation. Application of the quality criteria in the standard clusters "participation, independence and fairness" and "usability" was predominantly recorded at the level of the organisation – i.e. across all evaluations.

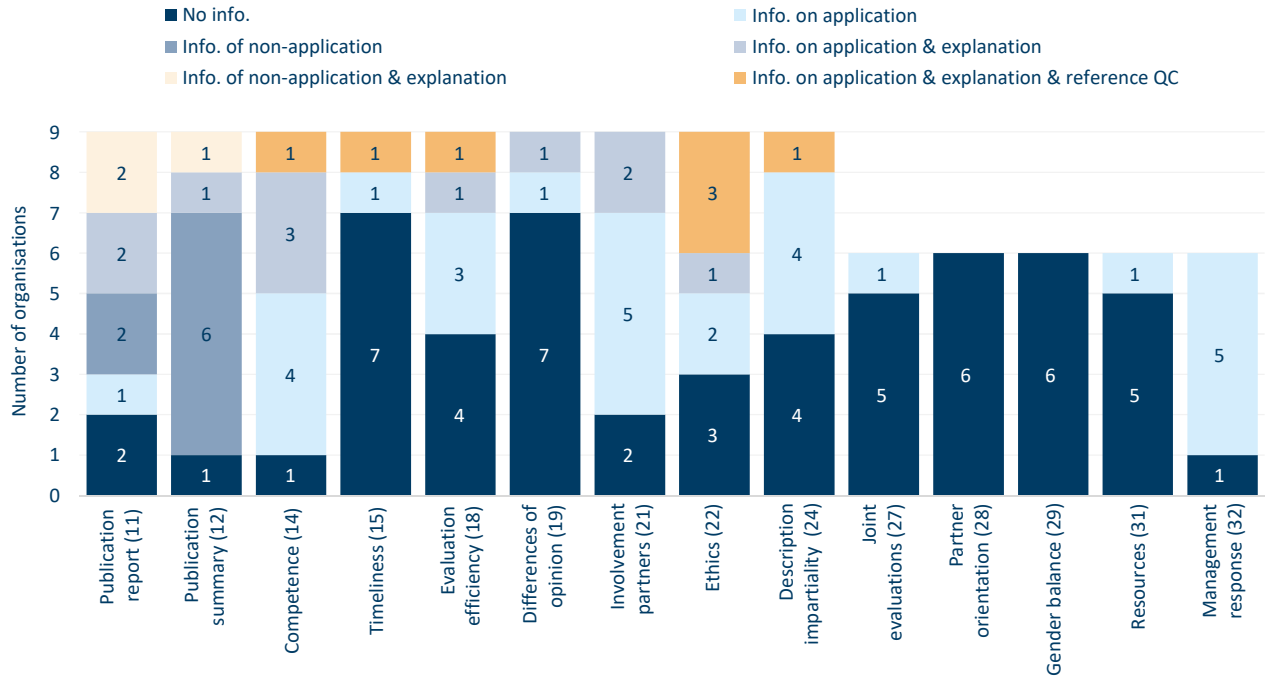
Barely any (explained) non-applications of a quality criterion were documented at the organisational level. Examples include the "publication of the evaluation report" and the "publication of the executive summary". At the evaluation level, there were a few explained non-applications of the "application of the OECD-DAC criteria". Non-application was prescribed and explained for two of the 14 quality criteria examined, which were therefore rated overall as "barely achieved"[63] (N = 2 organisations, for the quality

63 The DeGEval standards document (DeGEval, 2016: 29) states: "The partial or complete non-application of individual standards should always be documented and explained clearly and explicitly, for example in reporting. This makes it possible to assess the quality of the evaluation." Without written records, assessing the quality of evaluation of the OECD-DAC standards is only possible to a limited extent. Consequently, documentation is also required for this.

criterion "publication of the evaluation report [11]"; N = 1 organisation, for the quality criterion "publication of the executive summary [12]"). For these two quality criteria, cases of non-application were also described without explanation (N = 2 organisations, for the quality criterion "publication of the evaluation report [11]"; N = 6 organisations, for the quality criterion "publication of the executive summary [12]"; Figure 8).

In the organisational documents examined, hardly any reference was made to a particular quality standard. Specifically, for five out of 14 quality criteria, at least one organisation referred in its organisational documents to the application of quality standards (N = 1 organisation, for the quality criterion "competence of the evaluators [14]"; N = 1 organisation, for the quality criterion "timeliness of the findings [15]"; N = 1 organisation, for the quality criterion "evaluation efficiency [18]"; N = 3 organisations, for the quality criterion "evaluation ethics [22]"; N = 1 organisation, for the quality criterion "description of the impartiality of the evaluators". For "partner-country orientation [28]" and "gender balance in the evaluation team [29]", no information was available in the organisational documents of any organisation (Figure 8).

Figure 8 Documentation of the application/non-application of selected quality criteria in the organisational documents of group 1



Source: DEval, authors' own graph

Note: QC = quality criterion. The findings for group 2 can be found in section 4.1.1 of the online annex. Information on non-application & explanation & reference to QC was also examined, but could not be found for any quality criterion.

When directly asked in the online survey, respondents gave a variety of explanations for (partial) non-application of the quality criteria examined. For some of the quality criteria, the explanations given by the organisations were similar, for others they differed. Furthermore, the reported explanations for non-application were on different levels: 1) the organisational level (e.g. responsibility for application lay outside the evaluation units/desks or the quality criterion had no relevance to the organisation); 2) the level of the evaluation (e.g. the quality criterion had no relevance to the particular evaluation), and/or 3) the quality criteria were applied in a way that differed from the way in which the criteria were operationalised in the meta-evaluation (e.g. quality assurance was not ensured through an inception report "quality assurance with inception report [8]"). Details on the explanations for (partial) non-application can be found in section 4.1.1 of the online annex.

The following graphs and written sections have a uniform structure.

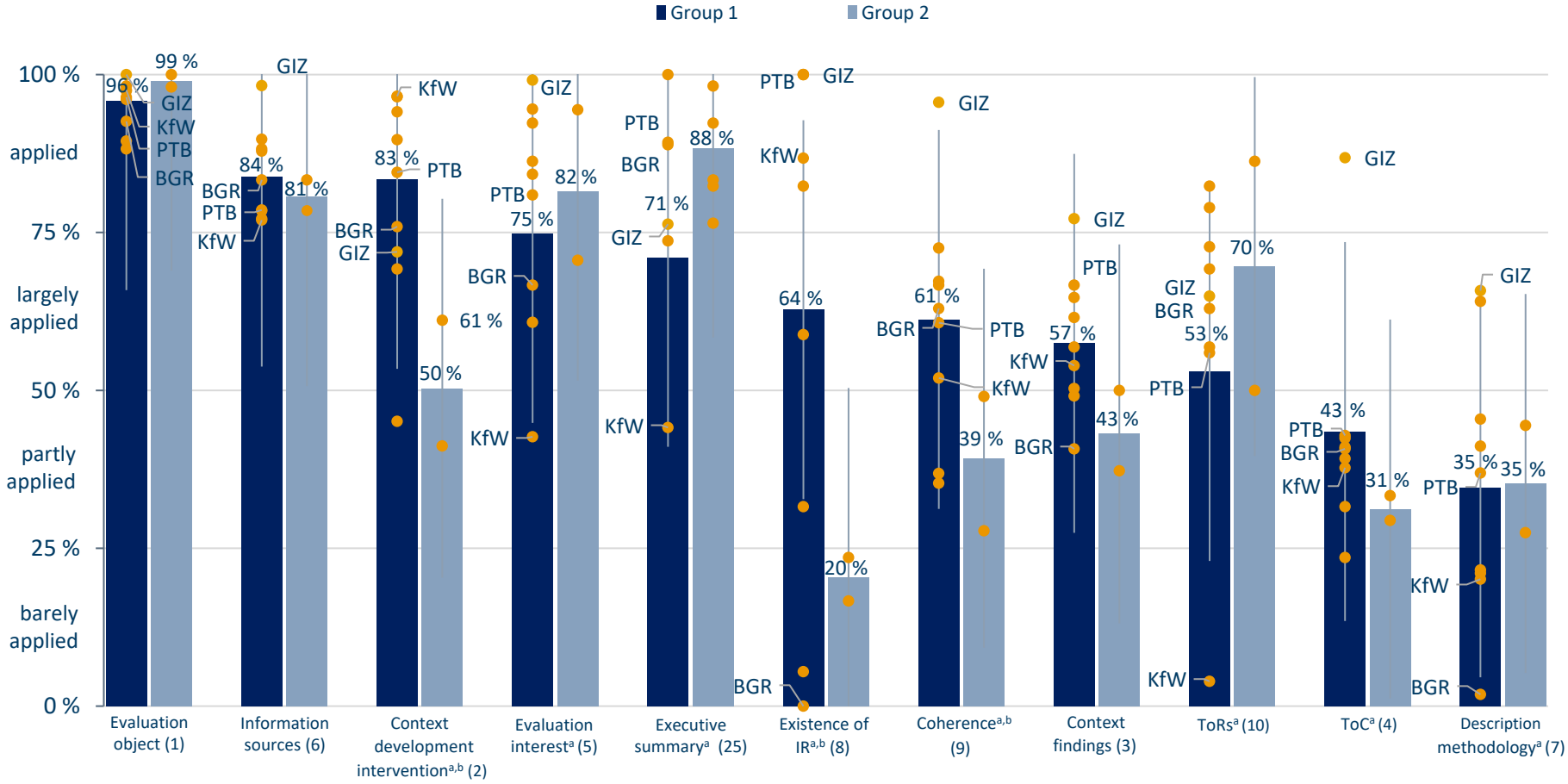
In the graphs on the average findings for the two groups and organisations (Figure 9, Figure 12, Figure 14 and Figure 16), the quality criteria of a standard cluster are arranged from left to right in descending order according to the average finding for group 1 (required to apply quality criteria). The dark blue/yellow bar in each quality criterion represents the average score as a percentage across the organisations in group 1; the light blue/yellow one does likewise for group 2 (not required to apply quality criteria). The individual average organisational findings are shown on the longitudinal axes of the group bars as yellow (Figure 9 and Figure 16) or blue (Figure 12 and Figure 14) dots. The percentage value is placed in the intervals of the y-axis, enabling the reader to read off the score. Due to the ambiguous mandate of DEval at the beginning of the meta-evaluation regarding the conduct of meta-evaluations of some non-governmental organisations, only the findings for the four official implementing organisations were identified by name in the graphs, in order to preserve anonymity. The thin black line on the longitudinal axis of all bars serves as a visual aid to enable the reader to clearly recognise to which group bar the yellow/blue dots belong. For the governmental organisations, the average findings are labelled; for the non-governmental organisations, they are shown anonymously.

The frequency graphs (Figure 10, Figure 13, Figure 15 and Figure 16) show the percentage shares for the two/four scores from the 296 evaluations for the quality criteria for each standard cluster. It is thus possible to see, for example, how many evaluations in a group have reached the highest score. For the standard clusters "participation, independence and fairness" and "usability", the order differs between the graphs for the average findings and the frequencies (i.e. between Figure 12, Figure 13, Figure 14 and Figure 15). This is because these standard clusters also include quality criteria from the online survey, whose scales differ from those at the evaluation level. Therefore, quality criteria at the organisational and evaluation levels are shown separately in the frequency graphs. However, the logic at both levels again follows the descending order in accordance with the finding for group 1.

The text sections describe the findings for each quality criterion (quality standard), and highlight in particular whether the organisations in group 1 within a quality criterion display a major difference between the minimum and the maximum organisational scores (> 50 per cent), or whether the average findings for the two groups differ in terms of absolute values by more than 20 per cent. All quality criteria are numbered uniformly in sections 4.2.1 and 4.2.2. An overview of the numbering can be found in section 7.1 of the annex to the report. The findings described in the text generally refer to group 1, and any exceptions are highlighted explicitly. When the "application" of a quality criterion is described with a score, the finding is descriptive. When the text refers to "achievement", this involves a rating. The text that follows the graphs is always structured according to the same arrangement of the quality criteria as in the graph. However, some text sections may deviate from the order shown in the graph, in cases where several quality criteria have been assigned to the same quality standard. In these cases, all associated quality criteria are dealt with directly in one section as soon as the first quality criterion in the graph is described. For some quality criteria, good practice examples are also presented.

Standard cluster "reporting and methods"

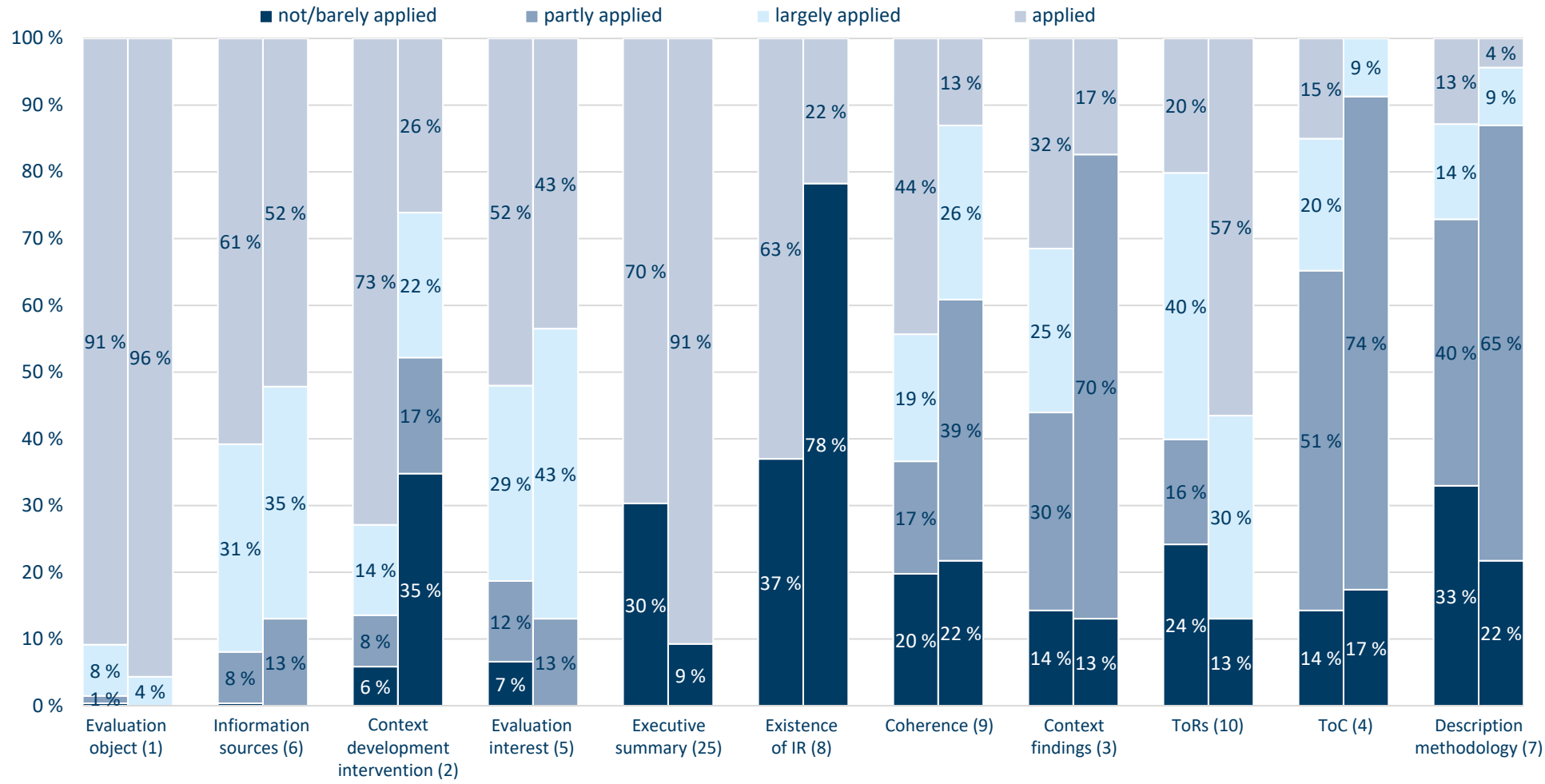
Figure 9 Application of the quality criteria in the standard cluster "reporting and methods"



Source: DEval, authors' own graph

Note: Some organisations applied quality assurance processes that differed from "quality assurance with inception report (8)" (e.g. "comments on evaluation reports"). As these processes were not examined in the present meta-evaluation, application by these organisations is underestimated. [a] Quality criteria show a difference between the minimum and maximum values for the organisations of > 50 per cent in group 1. [b] Quality criteria differ, with group 1 > group 2 (20 per cent).

Figure 10 Frequencies of the scores for the quality criteria in the standard cluster "reporting and methods"



Source: DEval, authors' own graph

Note: The left bar of each of the two bars shown next to each other = group 1; right bar = group 2. When adding up the percentages for each quality criterion, a deviation of +/- 1 per cent to 100 per cent may occur due to rounding. The numbering of the quality criteria supports cross-referencing with the text. An overview of the numbering can be found in section 7.1 of the annex to the report.

***Description of the evaluation object (1)*: On average, this quality criterion was "achieved". Thus, all organisations ensured almost throughout their evaluation reports that the development intervention under evaluation was comprehensibly described.** The deviation from a rating of 100 per cent means that in a few evaluations (9 per cent), objectives, target groups and/or relevant actors for the evaluated development evaluation were not described (Figure 10). In particular, the evaluation reports omitted to mention the target group or relevant actors in these cases. The GIZ exceeded the quality standard, i.e. all relevant aspects of the evaluation object were described in each evaluation. Its evaluation reports can be considered good practice in this respect. (For a good practice example of detailed description of the objectives, target groups and relevant actors of an evaluation, see section 4.1.1 of the online annex).

***Clarity of the information sources (6)*: On average, this quality criterion was "achieved". In about 92 per cent of the evaluation reports, the sources of information were at least largely described.** This means that the evaluation documents described transparently which specific documents or surveys were used as sources of information for which analysis. For the remaining 8 per cent, the sources of information were presented in a more rudimentary way (for example, it was only stated that interviews were conducted and secondary data were used). Within an evaluation, the different sources of information were presented in varying degrees of detail. In the analysis of possible factors affecting the application of this quality criterion, it became apparent that, among other things, the addition of different quality assurance processes (for example, "quality assurance with inception report [8]" and "stakeholder involvement [20]") was linked to clearer information sources.

***Consideration of the context (2 + 3)*: The two quality criteria for this quality standard were on average "achieved" and "largely achieved". Overall, the "description of the context of the development intervention (2)" was explained frequently, but was barely taken into account systematically in the findings obtained.** The quality criterion "description of the context of the development intervention" examined how many contextual elements (e.g. political, economic) were described, and how comprehensively, in the evaluation report and its annexes. The history of the organisation and its activities in the partner country were not considered as contextual elements. In particular, the involved political foundations, which per se operate in a political context, included a detailed and extensive contextualisation of the development intervention in many of their evaluation reports. The quality criterion "incorporation of the context in the findings (3)" showed how systematically contexts were considered in the findings (effectiveness). Only two organisations took the context into account by describing the factors that limited and those that enabled the findings. Good practice examples were identified for both quality criteria at the PTB, KfW and GIZ, among others. These showed that the quality criterion was achieved particularly well if a section was provided for the description of the context in the introduction, as well as in the description of the findings, as part of a standardised structure of the evaluation report (online appendix, section 4.1.1).

***Description of the evaluation interest (5):* On average, this quality criterion was "largely achieved".**[64] **Specifically, the purpose, objective and evaluation questions of the evaluation were clearly described in about 52 per cent of the evaluation reports/annexes.** The aspect that was most frequently not clearly explained in evaluation reports/annexes was the overarching purpose of the evaluation (for example, reviewing whether a development intervention should be extended). Evaluation questions were described in almost all evaluation reports, and the objectives of an evaluation (for example, assessing the effectiveness of a development intervention) were often described in evaluation reports. This quality criterion was achieved by MISEREOR, among others, because the use of a template for terms of reference (ToRs) meant that all three aspects of the evaluation interest had to be worded in a standardised way (online annex, section 4.1.1).

64 The "description of the evaluation interest (5)" is the only quality criterion for which application is rated higher in relation to the calculated median values (i.e. achieved rather than largely achieved) than the calculated mean values.

Information content of the executive summary (25): **On average, this quality criterion was "largely achieved". In about 70 per cent of the executive summaries, the findings and conclusions or recommendations of the evaluation were described.** The content of an executive summary was rated as "achieved" if the findings and either conclusions or recommendations of the evaluation were specified. The majority of organisations achieved this quality criterion, one organisation largely achieved it and one partly achieved it. Since an evaluation does not necessarily have to make recommendations, the conclusions were also used as the basis for the assessment, in accordance with the standards documents. In the regression analyses to explain the application of the quality criterion, it was found that the information content of the executive summary increased, the more evaluators were involved in the evaluation.

Quality assurance with inception report (8)**: On average, this quality criterion was "largely achieved". It focuses on the existence of an inception report, and thus on one form of quality assurance.** Based on a focus group discussion with the involved organisations and existing literature (Queiroz de Souza, 2017), the production of an inception report was identified as an important feature of quality assurance processes. An inception report is an effective tool for creating a common understanding of the procedure and implementation of the evaluation. It is also a good tool for discussing and, if necessary, adjusting any initial critical aspects. An inception report was available for around 63 per cent of the evaluations. The PTB and GIZ produced inception reports for all their evaluations, while other organisations did not produce any inception reports for any of their evaluations. Hence there is a wide variation in the application of this quality criterion. The online survey showed that, in addition to the inception report, organisations applied alternative forms of quality assurance process (including the use of templates for the terms of reference, or annotated outlines for the draft report; see section 4.1.1 of the online annex). Accordingly, it can be assumed that the actual implementation of evaluation quality assurance was not fully captured by the operationalisation of this quality criterion.

Coherence of data-findings-conclusions (9): **On average, this quality criterion was "largely achieved". In approximately 44 per cent of all evaluation reports, the majority of conclusions were coherently based on data analysis and findings.** In about 37 per cent of the evaluation reports, for the majority of the conclusions it was barely or only partially possible to trace these back to the findings and/or data analyses on which they were based. GIZ was the only organisation to fully apply this quality criterion. As a good practice example from the PTB evaluations, it turned out that the quality criterion was well achieved when direct references to the findings were included in brackets (for example, using page numbers and/or footnotes) when writing up the conclusions. Another possibility is framing the conclusions in the same paragraph (see section 4.1.1 of the online annex for good practice examples).

Information content of the terms of reference (10): **On average, this quality criterion was "largely achieved". In about 60 per cent of the terms of reference, at least four out of eight aspects were specified (purpose, beneficiaries, objectives, methods, time frame, available resources, publication rights and persons involved in the evaluation).** The publication rights as well as the naming of the beneficiaries and the persons involved in the evaluation were described the least. For one organisation, the quality criterion was rated as "barely achieved". One reason for this was that a terms of reference was not necessarily part of an evaluation for this organisation.

Description of the Theory of Change: **On average, this quality criterion was "party achieved". Although almost all evaluations worked with elements of a Theory of Change (input - output - outcome - impact), these were often incompletely illustrated or not illustrated for all objectives of a development intervention.** A full description of the Theory of Change facilitates understanding of how the development intervention works, and can be used as a basis for impact-oriented evaluations (UNDAF, 2017). In 14 per cent of the evaluations, no Theories of Change were described. In 71 per cent incomplete Theories of Change were described (in 51 per cent of cases, incomplete Theories of Change were formulated and in 20 per cent, complete Theories of Change were formulated for at least one objective of the development intervention, but not for all objectives). In 15 per cent of evaluations, complete Theories of Change were developed for all objectives of the development intervention. These came predominantly from GIZ evaluations (upward outliers). Some organisations explained their low application with a lack of available resources or a lack of commitment by the commissioner. It is also possible that the Theory of Change was described in documents

that were not made available for the meta-evaluation. Suitable good practice examples are evaluations of the GIZ, in which the Theory of Change was presented in a separate section using a graphic plus a written explanation of the causal relationships (see section 4.1.1 of the online annex for good practice examples).

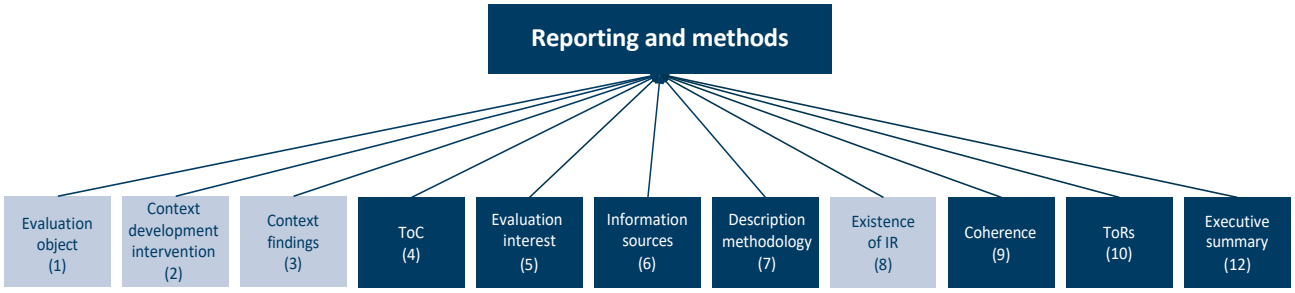
***Description of the methodological adequacy (7)*: On average, this quality criterion was "partly achieved". In the majority of the evaluation reports, limitations of the approach were described. Why the methodological approach had been chosen was rarely explained, and there was almost never a discussion of alternative approaches.** This quality criterion examined whether or to what extent a logical explanation was given as to why the applied methods were appropriate, and whether or to what extent limitations of the methodological approach were discussed. No organisation applied the quality criterion fully, and four explained their approach on average barely at all. The explanations given constitute key information that would enable users of the evaluation to assess the reliability and validity of the findings. A total of 13 per cent of the evaluation reports provided comprehensive explanations of why the chosen methods were appropriate for the evaluation and what limitations they produced.[65] In 14 per cent of cases, both aspects were described in a few sentences, though not comprehensively, and in 40 per cent either explanations of why the approach was chosen or limitations were listed. In 33 per cent of the evaluations, neither explanations why nor limitations were discussed. Taken together, more than 73 per cent of the evaluations thus barely or partly achieved this quality criterion. There are great differences in the findings between the organisations. Good practice examples were identified in the evaluations of the DVV, in which detailed sections were exclusively dedicated to the advantages and disadvantages as well as the limitations of the methods used (see section 4.1.1 of the online annex for good practice examples).

Finally, the findings of an empirical investigation showed that most of the quality criteria of the standard cluster "reporting and methods" could be related to one factor (the standard cluster). The quality criteria were assigned to standard clusters. Whether these standard clusters reflect commonalities not only theoretically but also empirically, was examined by means of an exploratory factor analysis. The analysis showed that seven quality criteria could be empirically represented by the standard cluster "reporting and methods": "description of the Theory of Change (4)", "description of the evaluation interest (5)", "clarity of the information sources (6)", "description of the methodological adequacy (7)", "coherence of data-findings-conclusions (9)", "information content of the terms of reference (10)" and "information content of the executive summary (25)". Four quality criteria are not represented empirically ("description of the evaluation object [1]", "description of the context of the development intervention [2]", "incorporation of the context in the findings [3]" and "quality assurance with inception report [8]"; Figure 11).[66]

65 Text mining was also used to investigate whether guided interviews or focus group discussions were used as data collection methods in the evaluations analysed. The findings of this analysis show that guided interviews were used in 98 per cent of the evaluations (N = 291) and focus group discussions in 47 per cent of the evaluations (N = 138). For more information on the text mining method and results, see section 4.1.2 of the online appendix.

66 The quality standards "description of the evaluation object (1)" and "consideration of the context" might not be represented by the standard cluster as they reflect information on the development intervention and not on the evaluation. They represent a separate factor in their own right. For more detailed information on the factor analysis, see section 3.4 of the online annex.

Figure 11 Exploratory factor analysis of the standard cluster "reporting and methods"

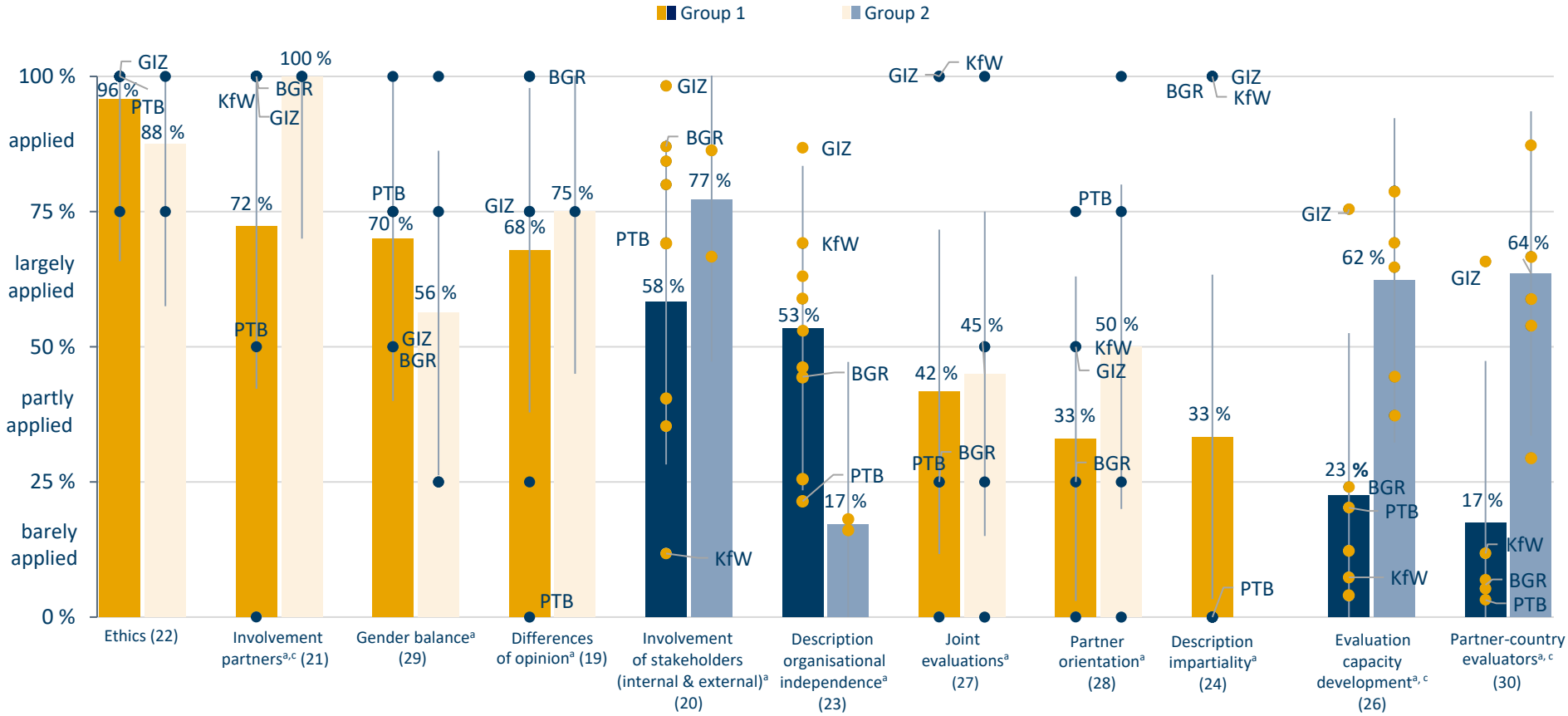


Source: DEval, authors' own graphic

Note: The arrows represent factor loadings, and indicate that the analysis examined whether the quality criterion is linked to the standard cluster. The quality criteria highlighted in grey are not part of the construct/standard cluster identified by the exploratory factor analysis.

Standard cluster "participation, independence and fairness"

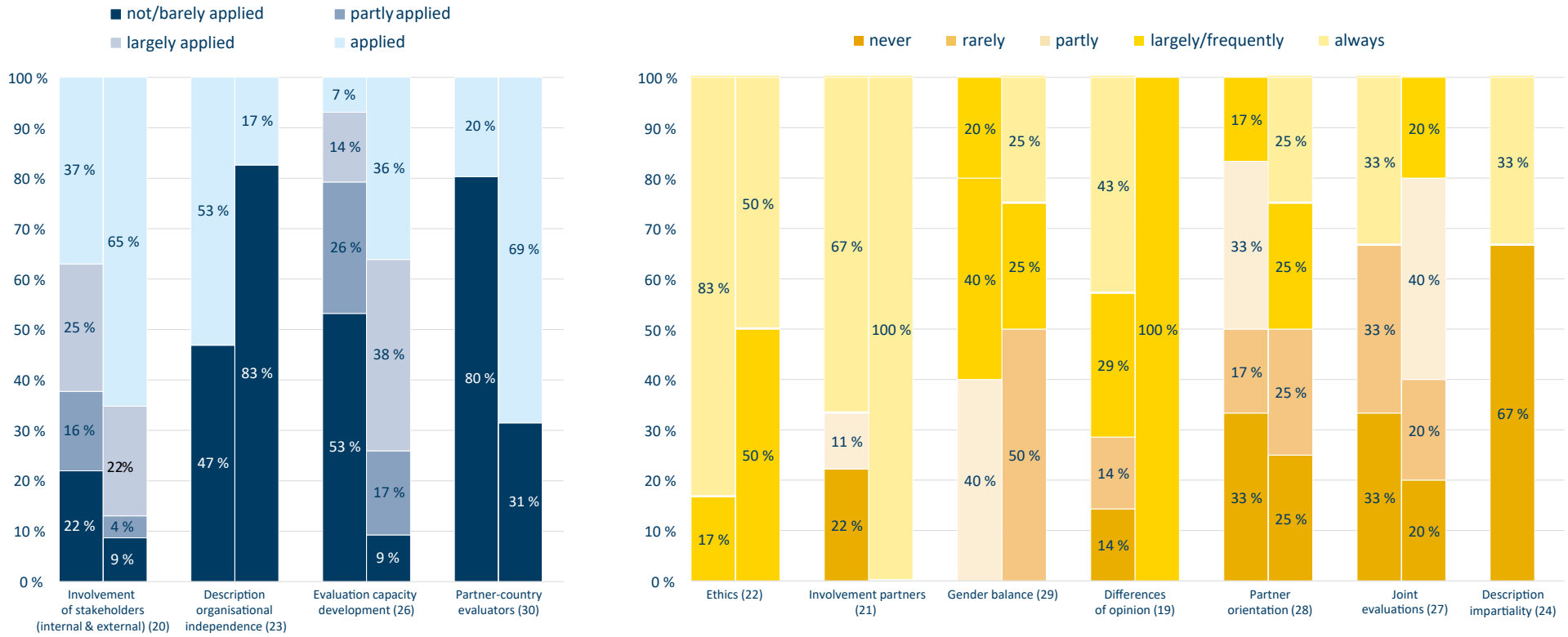
Figure 12 Application of the quality criteria in the standard cluster "participation, independence and fairness"



Source: DEval, authors' own graph

Note: blue = quality criteria (QCs) examined at the level of individual evaluations; yellow = QCs examined at the level of the organisation. For the quality criterion "description of the impartiality of the evaluators (24)", organisations in group 2 did not provide any information in the online survey, therefore the light-coloured bar was omitted. Some organisations applied this quality criterion in other ways (for example, in the context of selection processes rather than through a signed declaration of impartiality). As these ways of applying the criterion were not examined in this meta-evaluation, these organisations are underestimated in terms of their application of it. [a] Quality criteria show a > 50 per cent difference between the minimum and maximum values for the organisations. [b] Quality criteria differ, with group 1 > group 2 (by 20 per cent). [c] Quality criteria, differ with group 1 < group 2 (by 20 per cent). An overview of the numbering can be found in section 7.1 of the annex to the report.

Figure 13 Frequencies of the quality criteria scores in the standard cluster "participation, independence and fairness"



Source: DEval, authors' own graph

Note: The left bar of each of the two bars shown next to each other = group 1; right bar = group 2; blue = quality criteria (QCs) examined at the level of individual evaluations; yellow = QCs examined at the level of the organisation. When adding up the percentages for each quality criterion, a deviation of +/- 1 per cent to 100 per cent may occur due to rounding. For the quality criterion "description of the impartiality of the evaluators (24)", organisations in group 2 did not provide any information, which is why only the bar for group 1 is shown. The arrangement of the quality criteria may differ from the arrangement in the previous graph (see also explanation of the graphs in subsection 4.2.1). The numbering of the quality criteria supports cross-referencing with the text. For an overview of the numbering, see section 7.1 of the annex to the report.

***Evaluation ethics (22)*: On average, this quality criterion was "achieved". All organisations reported online that they always, or at least frequently, included guidelines on protection and the rights of evaluation participants in the evaluations.** The quality criterion was not narrowed down. This meant that its application could include (i) a variety of content (such as data protection, human rights or protection of persons or groups of persons interviewed in fragile contexts), and (ii) alternative forms of implementation (for example, involving a code of conduct signed separately by evaluators, or through training on ethical conduct in evaluations). In six out of nine organisations, guidelines for application of the quality criterion were included in organisational documents (e.g. evaluation guidelines or contracts between evaluators and organisations). In three cases a reference to the quality standards was included.

***Involvement of internal and external stakeholders (20) and partners (21):* On average, both these quality criteria were "largely achieved". It emerged that stakeholders and partners were on average included more by the involved organisations that were not required to apply quality standards.** The first quality criterion describes the inclusion of at least one stakeholder in different evaluation phases. Internal stakeholders are, for example, the commissioners, and external stakeholders are the partner organisations in the partner country.[67] On average, group 2 achieved this quality criterion, and thus performed better than group 1. Furthermore, on average the organisations rated the quality criterion "involvement of partners (21)" in the online survey as "largely applied".[68] Explanations given for not including partners were, for example, that it was not possible to include them due to the excessive coordination work this would entail, or due to political sensitivities. The latter applies to the involved political foundations, among others. This is why an agreement between the BMZ and the political foundations states that evaluation principles such as the participation of partners can only be applied to a limited extent (BMZ, 2016).[69]

However, the measurement of this quality criterion has its limits. It is difficult to verify in what form (e.g. free expression of views) and with what intensity the inclusion of stakeholders was ensured in the overall process or in individual evaluation phases. A more detailed operationalisation and a closer analysis of this quality criterion would be necessary in order to obtain more in-depth information. One critical point to note is that the involvement of only one stakeholder in different evaluation phases can also be problematic (for example, if this stakeholder reflects a one-sided perspective on the findings). As an example of good practice, GIZ evaluations have shown that the quality criterion was achieved when the annex to the evaluation report contained a timetable with an additional column indicating whether and to what extent stakeholders were involved in the evaluation during the design, implementation and reporting phases. In the design phase, this was often achieved by involving stakeholders in generating the evaluation questions, or by working out in joint workshops which methods should be used to collect which information. In the reporting phase, however, stakeholders had the opportunity to comment on the draft report or to discuss the findings and recommendations before they were published (see section 4.1.1 of the online annex for good practice examples).

***Composition of the evaluation team (29 + 30):* In the case of official implementing organisations and the involved political foundations, there was on average barely any "inclusion of partner-country experts (30)". By contrast, non-governmental organisations in group 2 largely included them in the expert teams. A "gender balance in the evaluation team (29)" was largely considered in both groups, but not formally**

67 Persons, groups of persons or organisations that have something to lose or to gain are regarded as stakeholders in the evaluation. They can be people who are responsible for planning and implementing the evaluation: commissioners; funders; people responsible for the project being evaluated; people who wish to use/could use the evaluation (after Beywl and Niestroj, 2009).

68 The quality criterion does not distinguish between internal and external stakeholders. De facto, mainly internal stakeholders were coded in the evaluation documents. Yet the wording of the quality standard leaves room for interpretation, and it may mean exclusively external stakeholders. Consequently, additional information on the "involvement of partners" was collected in the online survey. The two quality criteria are therefore described and analysed together.

69 Since the wording in the agreement is not clear, the political foundations were analysed according to their application of the quality criterion (i.e. the documentation was not interpreted as implying an explained non-application).

recorded in organisational documents.[70] Overall, evaluators from partner countries were involved in around 20 per cent of group 1 evaluations, and in around 70 per cent of group 2 evaluations. The GIZ largely applied the quality criterion, and thus represents a positive exception among the organisations required to apply quality standards. The non-inclusion of partner-country experts was often explained by the fact that only one person rather than a team of evaluators was hired. The PTB also applied the quality criterion in a different way. It hired experts from neighbouring countries of the partner country, in order to avoid potential bias due to personal links with the partner organisations in the national context. Challenges in measuring the quality criterion included the fact that partner-country evaluators can be involved in evaluations of development interventions in several countries, and that it was difficult to identify the evaluators' affiliation to the partner country. It is therefore difficult to meet this quality criterion in evaluations of development interventions that involve several countries, and the finding for the quality criterion is to be assessed conservatively. With regard to the consideration of a "gender balance in the evaluation team (29)", some organisations stated that they sought to achieve a gender balance not for individual evaluations, but across evaluations. In the organisational documents, two organisations made reference to the consideration of a balanced male-female ratio.

***Transparency of differences of opinion (19)*: On average, this quality criterion was "largely achieved". Making differences of opinion between members of the evaluation team transparent in the evaluation report did not seem desirable to all organisations.** Some organisations stated in the online survey that the reason for not applying this criterion was that rather than describing differences, their goal was consensus among the evaluators. Overall, the quality criterion was difficult to capture at the level of the individual evaluation, as in most cases it was not clear whether there were differences of opinion among the evaluators.

***Independence of the evaluators (23 + 24):* This quality criterion incorporated the "description of the organisational independence of the evaluators (23)" ("largely achieved" on average) and the "description of the impartiality of the evaluators (24)" ("partly achieved" on average).** In 53 per cent of the evaluations, the evaluators were not involved in the development intervention either politically or operationally or in an advisory capacity, and did not belong to its target group. In 47 per cent of cases they were organisationally dependent. In these cases, however, 40 per cent of the evaluations did not permit any conclusions to be drawn about independence, and the evaluators were therefore classified as not independent – as their independence was not documented, but could have been described. Regarding the personal "description of [the] impartiality of the evaluators (24)" vis-à-vis the evaluation object, three organisations (33 per cent) stated that they had this formally confirmed through a signed "declaration of impartiality". The remaining organisations declared that they had applied the guarantee of impartiality in other ways (for example, they had explained this in annexes to the contract). Overall, there were limitations in the measurement of this quality criterion too. For example, it would have been difficult to determine whether persons had worked for the evaluated development intervention before they became evaluators, and might thus have been biased towards their work. Since both quality criteria were estimated rather conservatively[71], the finding for the overarching quality standard was also interpreted conservatively.

***Consideration of joint evaluations (27):* The consideration of a possible joint evaluation with donors and/or the partner government in a partner country was "partly achieved" by the organisations.** One of the explanations given by organisations for not applying this criterion was that joint evaluations were not possible due to complex cooperation relationships involving different focal areas. A further reason was that the effort required for a joint evaluation would have been disproportionate to the size of the individual development interventions and/or the evaluation. Since non-governmental organisations rarely cooperate with partner governments in their development interventions, for them this quality criterion tended to be underestimated.

70 The two quality criteria were derived from the OECD-DAC Standard 3.1 "Evaluation team": "Gender balance is considered and the team includes professionals from the partner countries or regions concerned".

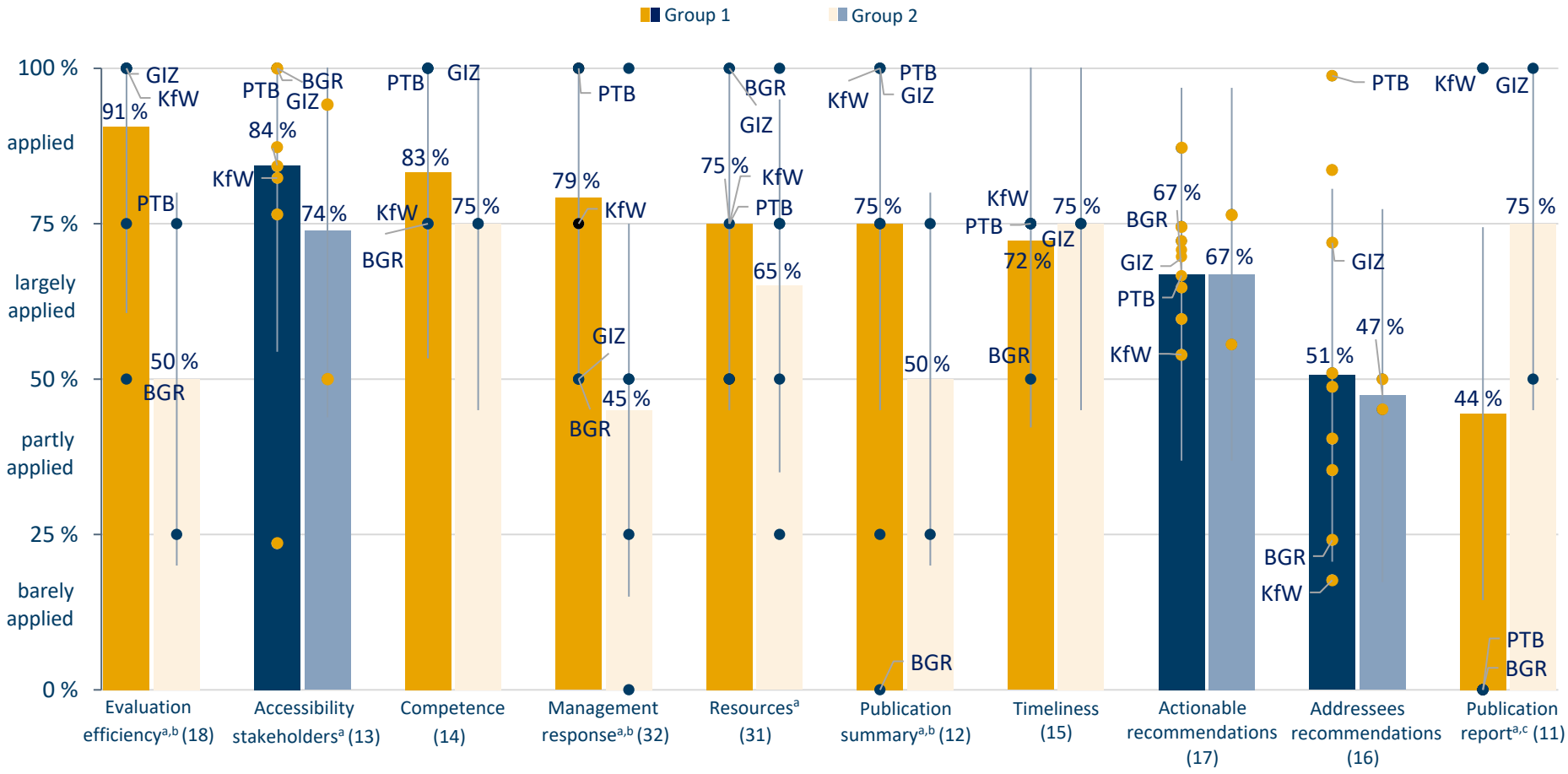
71 "Conservatively" means that the finding represents the lower value, and the true value is probably higher.

***Partner-country orientation (28)*: On average, this quality criterion was "partly achieved". The majority of organisations stated that "local or national evaluation activities and plans" were not systematically considered in the evaluations.** Here, partner orientation is understood as national evaluation guidelines or standards, for example. The findings show a wide spread between the organisations. In the online survey, the organisations explained non-application by stating that including such approaches was often not relevant or that their own development interventions were often too specific to establish a link. In the organisational documents too, there were also almost no references to the partner-country orientation. As with the quality criteria "involvement of partners (21)", "incorporation of evaluation capacity development (26)" and "inclusion of partner-country evaluators (30)", group 2 applied the quality standard more frequently in absolute figures.

***Incorporation of evaluation capacity development (26)*: On average, this quality criterion was "barely achieved" by the official implementing organisations and the involved political foundations. It thus represents one of the least applied quality criteria. By contrast, the remaining non-governmental organisations on average largely applied it.** Evaluation capacity development was not achieved in about 53 per cent of the evaluations. This shows that there was no focus on this quality standard in group 1. A recommendation by the OECD-DAC in its peer review (2021: 7) runs along similar lines: "Germany should continue to invest in building evaluation capacities in its partner countries and invest more in learning from evaluations of special initiatives and its overall investments at country, regional and programme levels." Organisations not required to apply quality standards actively integrated local/national evaluators and partner organisations in the design and implementation of the evaluation (for example by jointly developing evaluation questions), and discussed the evaluation findings with them more often than organisations that were required to apply quality standards, with the exception of GIZ. Since non-governmental organisations rarely cooperated with partner governments in their development interventions, they largely did not achieve one out of four aspects of evaluation capacity development, and the quality criterion tended to be underestimated for them. As with the independence of the evaluators, it was not possible to draw any conclusions about the application of this quality criterion in many evaluations. Hence the rating should be interpreted rather conservatively. As a good practice example, the timeline in GIZ evaluations can again be cited. This shows that partners and relevant ministries were involved at various stages in activities that promoted evaluation capacity development (see section 4.1.1 of the online annex for good practice examples).

Standard cluster "usability"

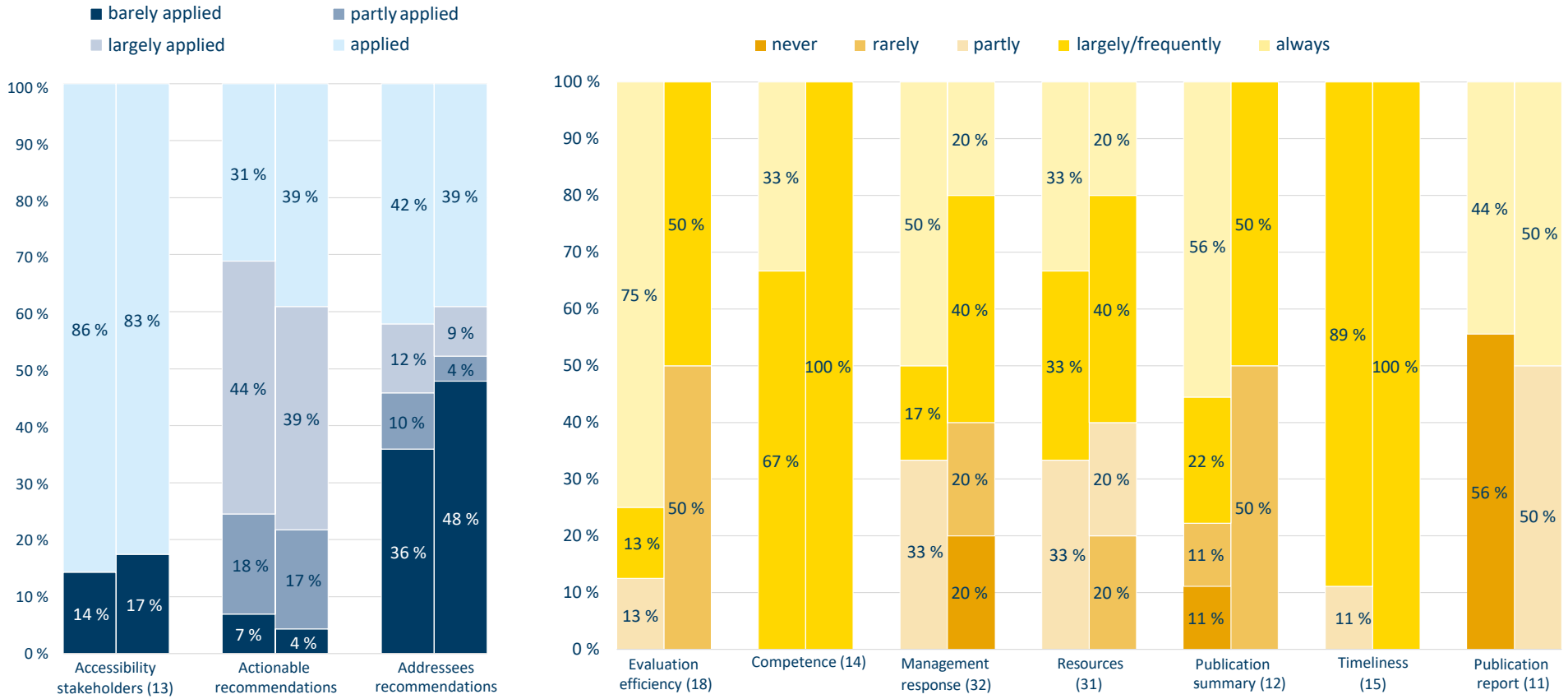
Figure 14 Application of the quality criteria in the standard cluster "usability"



Source: DEval, authors' own graph

Note: blue = quality criteria (QCs) examined at the level of individual evaluations; yellow = QCs examined at the level of the organisation. [a] Quality criteria show a difference between the minimum and maximum values for the organisations of > 50 per cent in group 1. [b] Quality criteria differ with group 1 > group 2 (20 per cent). [c] Quality criteria differ, with group 1 < group 2 (by 20 per cent).

Figure 15 Frequencies of the scores for the quality criteria in the standard cluster "usability"



Source: DEval, authors' own graph

Note: The left bar of each of the two bars shown next to each other = group 1; right bar = group 2; blue = quality criteria (QCs) examined at the level of individual evaluations; yellow = QCs examined at the level of the organisation. When adding up the percentages for each quality criterion, a deviation of +/- 1 per cent to 100 per cent may occur due to rounding. The arrangement of the quality criteria may differ from the arrangement in the previous graph (see also explanation of the findings graphs in subsection 4.2.1). The numbering of the quality criteria supports cross-referencing with the text. For an overview of the numbering, see section 7.1 of the annex to the report.

Evaluation efficiency (18): **On average, this quality criterion was "achieved". Almost all organisations stated that they had – either always or frequently – reflected on the costs of evaluations in relation to their benefits.** At the GIZ, KfW and PTB, evaluation efficiency was considered by not evaluating all development interventions through representative sampling (at the PTB since 2020, previously full population analysis). Furthermore, these organisations mentioned that if the expected benefits were low, a desk study with telephone interviews would be conducted instead of on-site missions. Other organisations stated that evaluations were always useful for project management and the design of further projects, and thus always proportionate to their costs. The latter may indicate that evaluation effectiveness is considered mainly on the basis of benefits, rather than on the costs or the ratio between the two. This quality criterion could include reflection on the cost-benefit ratio before an evaluation (when considering whether to conduct it), and during an evaluation (when considering how implementation decisions relate to the expected benefits). However, the responses predominantly related to consideration of the ratio before the evaluation started. This indicates that there is no common understanding of evaluation effectiveness between organisations. It also means that in relevant cases, there was rarely any readjustment during the course of the evaluation.

Accessibility (11 + 12 + 13)**: The accessibility of the evaluation findings was examined for three ways of applying the standard: 1) accessibility of the evaluation findings for stakeholders (for example in the form of a joint discussion), 2) accessibility of the executive summary for the public and 3) accessibility of the evaluation report for the public. On average, the quality standard as a whole was "largely achieved", although in individual cases this was due to an explained non-application. There were also clear differences in the three ways of applying the standard.** On average, the findings were made available to stakeholders in full, the evaluation reports were published in part and the summaries in large part. Due to the fact that non-application was explained in the organisational documents of three organisations, the quality criterion "publication of the evaluation report (11)" was rated as "achieved" on average.[72] This also applies to one organisation for the quality criterion "publication of the executive summary (12)". In summary, in about 86 per cent of the evaluations the evaluation was made available to stakeholders in the form of the final evaluation report, a written summary or a final presentation. Most organisations stated that they refrained from publishing the full evaluation reports in order to protect evaluation participants and partners. This is now required by the BMZ Evaluation Policy.[73] One organisation refrained from publishing sensitive evaluations. Another organisation wrote two separate parts of the report, and published only the one part that did not contain any sensitive information. Furthermore, in a BMZ agreement, the political foundations were granted leeway to pursue a reserved approach to the OECD-DAC evaluation principle of transparency (BMZ, 2016). This was because the political circumstances of some countries, for example, create a special need for protection of partner structures. A summary was published by the organisations on a website more often than the full evaluation report. The analysis of the organisational documents showed that accessibility was the only quality standard for which non-application was explained.

Competence of the evaluators (14): **In all organisations, the teams of evaluators had, on average, full evaluation, technical/sectoral and regional/country expertise.** It is evident that organisations prioritised different competences when selecting the team of evaluators. For example, they preferred (when required) technical/sectoral or evaluation knowledge over country knowledge, and sometimes included regional expertise rather than country expertise. It was also reported that attempts were made to compensate for weaknesses in evaluation expertise through close supervision by the responsible officers. One point of criticism to note is that the competence was assessed by the responsible officers in the evaluation units/desks for all evaluations of the respective organisation. This does not necessarily reflect the actual competence of the evaluators for each individual evaluation. Moreover, it is difficult to examine this quality

[72] An explained non-application should not be considered qualitatively equivalent to actual application. It should be logical and appropriately tailored to the context of the organisation concerned.

[73] In the adopted BMZ Evaluation Policy (BMZ, 2021a: 21), the quality criterion "publication of the evaluation report (11)" is highlighted as follows: "In the interests of transparency, reports should preferably be published in full".

criterion by obtaining objective data, as this involves confidential information on the evaluators (for example, a CV).

Existence of a management response (32): **On average, this quality criterion was "achieved" by group 1 and "partly achieved" by group 2. Not every organisation had an established process for writing a management response or for planning the implementation of recommendations.** As a management response is written after the evaluation report, its existence could only be verified through additional documents or the online survey. Three of the six organisations in group 1 always produced a management response, one did so frequently and two did so in some cases. Some organisations reported that they used other forms of response. For example, this included integrating implementation guidelines into the follow-up design of the evaluated project right away, or discussing them with partners at the end of the evaluation. In five of the six organisations, the quality criterion was specified in the organisational documents, though without reference to the standards documents (Figure 8). The BMZ Evaluation Policy that recently came into force requires at least a follow-up of the implementation of the recommendations (BMZ, 2021a).

Sufficient resources available (31): **According to officers responsible for evaluations, "financial resources, time and human resources" were on average largely sufficient to achieve the objectives of the evaluations.** The implementing organisations in group 1 are almost exclusively funded by taxpayers' money and not by other resources. These organisations reported that resources were either always or often sufficient to achieve the objectives. Group 2, which included only non-governmental organisations, shows a spread between organisational findings. On average, these organisations reported that their resources were largely sufficient. In some cases in group 2, funding for evaluations was supplemented by donations, or the performance-based component of the evaluators was reduced when internal budget reallocations were not possible.

Timeliness of the findings (15): **On average, organisations reported that their evaluations were largely completed on time by the agreed date.** When this was not the case, organisations attributed this either to external factors (e.g. the security situation in the partner country or the COVID-19 pandemic), and/or to internal factors (e.g. coordination loops between evaluators or availability of partners). In group 1, the timeliness of the completion of the evaluations was only addressed in two out of nine organisational documents (Figure 8).

Usefulness of the recommendations (16 + 17)**: On average, this quality standard – comprising the two quality criteria 1) "addressees of the recommendations (16)" and 2) "actionable recommendations (17)" – was "largely achieved".** Around 36 per cent of the evaluations named addressees (e.g. the partners of the development intervention or the evaluation unit of the organisations themselves) in fewer than a quarter of their recommendations. For 31 per cent, the implementation of more than half of the recommendations was a logical consequence of them. For example, the recommendations were considered actionable if they were worded as concrete instructions for implementation, making clear when and how a recommendation should be implemented. In contrast to the quality criterion "actionable recommendations (17)", "addressees of recommendations (16)" showed a wide spread between the organisational findings. In KfW evaluations, conclusions were assessed instead of recommendations.

4.2.2 OECD-DAC criteria

This second sub-section describes the findings on application of the five OECD-DAC criteria. As all eleven organisations were required to apply the OECD-DAC criteria, no group 2 is shown. Box 5 provides an overview of the general conclusion of the section.

Box 5 General conclusion on application of the OECD-DAC criteria

To what extent are strengths and weaknesses evident in the application of the OECD-DAC and the DeGEval standards in the evaluations of the involved German development cooperation organisations? (Evaluation question 2a)

The OECD-DAC criteria were very largely achieved by the involved German development cooperation organisations. However, it should be explicitly pointed out that the criteria were operationalised by examining individual questions in accordance with the BMZ guidelines (2006), rather than the comprehensive content of the OECD-DAC criteria. This made it easy to achieve the benchmark. In the application of the OECD-DAC criteria, first documentation of non-application was also available at the organisational and evaluation levels. In this respect, the application of the OECD-DAC criteria already differed – albeit to a minor extent – from the application of most of the other quality criteria. It can be assumed that in future evaluations, documentation of the non-application of the OECD-DAC criteria will continue to increase. This is because since 2020 (BMZ, 2020), the updated BMZ guidelines on evaluation criteria require priority setting that is explained and transparent.

a) Strengths were evident in the application of the OECD-DAC quality standards. (Finding 6)

- On average, the OECD-DAC criteria were achieved by more than 95 per cent. (Finding 6.1)
- There were first cases documenting the explanation of non-application. Worth mentioning here above all are the BMZ funding guidelines for the political foundations, which take into account the foundations' special circumstances when dealing with the OECD-DAC criteria. (Finding 6.2)

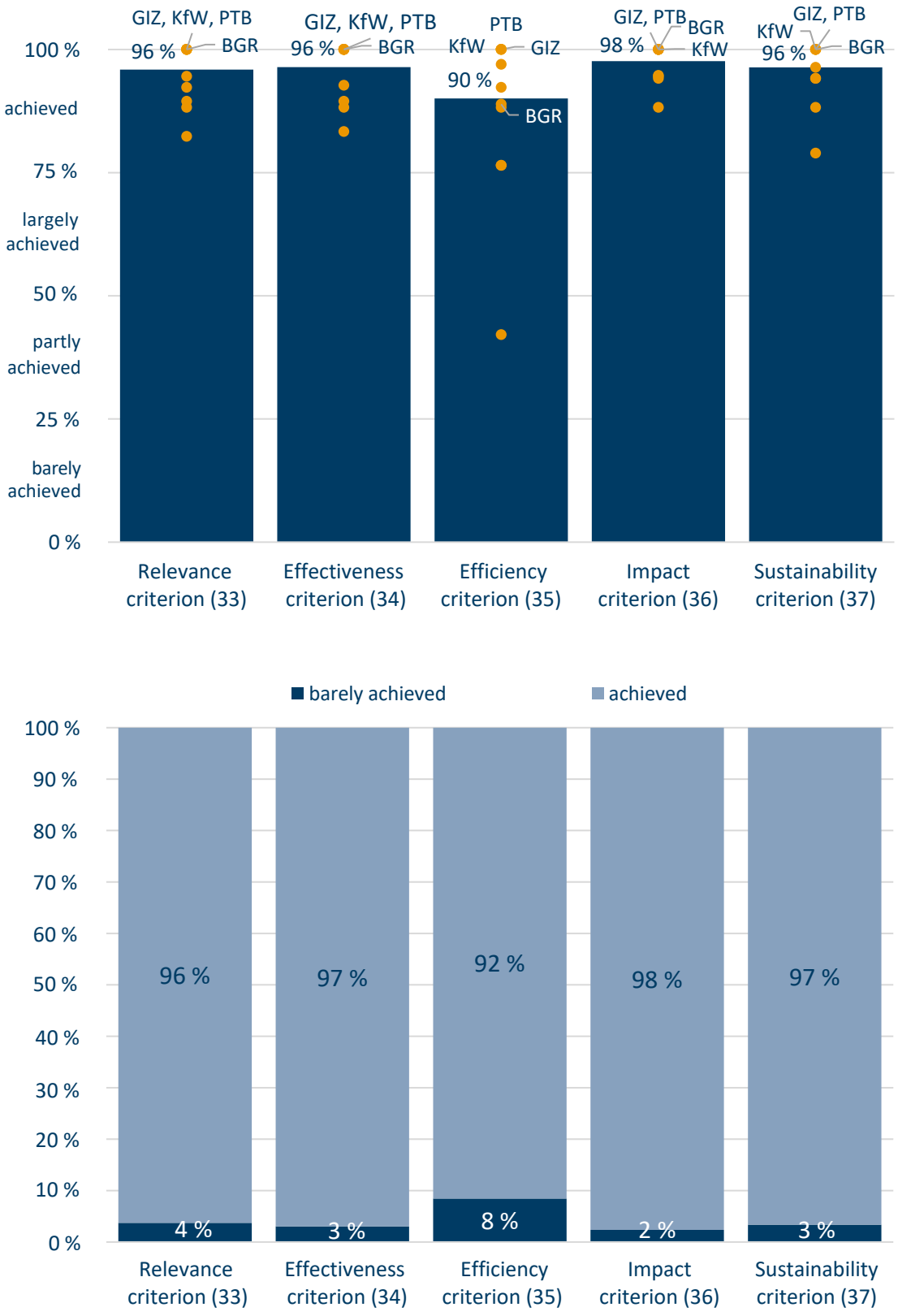
b) Weaknesses were found in the documentation and explanation of non-application. (Finding 7)

- A (partial) non-application of the quality criteria was not always clearly documented and explained. (Finding 7.1)

Application of the OECD-DAC criteria: **On average, all OECD-DAC criteria were "achieved" across all eleven organisations.** For the OECD-DAC criteria, the meta-evaluation established whether in the evaluations at least one question of the BMZ guidelines (BMZ, 2006) was adequately addressed in relation to "relevance (33)", "effectiveness (34)", "efficiency (35)", "impact (36)" and "sustainability (37)".[74] The governmental organisations achieved all the criteria, and in several cases exceeded the benchmarks (Figure 16). One organisation applied the OECD-DAC criterion "efficiency" partially, and referred in evaluations in some cases to a provision in the BMZ guidelines that allows political foundations to deal with all criteria more restrictively (BMZ, 2016). For all criteria, an explanation of non-application was rated positively. However, the findings almost always reflect achievement of the criterion through actual application, and only rarely achievement of the criterion through explained non-application. The estimated proportion of evaluations in which a criterion was positively rated based on an explained non-application is lower than 5 per cent.

[74] This analysis of the appropriate application of the OECD-DAC criteria was derived from the OECD-DAC standards, and not from the OECD-DAC publication on the criteria. The examination was carried out on the basis of one question from the BMZ guidelines (2006). The criterion "coherence", which has been required for organisations since 2020 (BMZ, 2020), was not included in the meta-evaluation due to the chosen time frame.

Figure 16 Achievement and frequencies of the ratings in the area "OECD-DAC criteria"



Source: DEval, authors' own graph

Note: The dark blue bars on the upper graph represent the average findings across all organisations. The lower graph with predominantly light blue bars shows the distribution of frequencies across all reports. No two groups exist for the graph, as all organisations were required to apply the OECD-DAC criteria.

4.2.3 Organisation-specific criteria

This third sub-section describes the findings on the application of the eleven organisation-specific quality criteria of four organisations. Organisation-specific quality criteria focused on subject matter that was important to an organisation in addition to the OECD-DAC and DeGEval standards, and should be included in every evaluation. Box 6 shows the general conclusion and the findings of the analysis.

Box 6 General conclusion on the application of organisation-specific quality standards

To what extent are strengths and weaknesses evident in the application of the organisation-specific quality standards in the evaluations of the involved German development cooperation organisations? (Evaluation question 2b)

A positive picture also emerged for the application of the organisation-specific quality standards by DRK, EWDE, GIZ and hbs. On average, these quality criteria were "largely achieved". Once again, there was potential for improvement in the explanation of non-application at the evaluation level.

a) Strengths were evident in the application of the quality standards. (Finding 8)

- On average, all organisation-specific quality criteria were "largely achieved" (71 per cent) – which is slightly lower than for the required OECD-DAC and DeGEval quality criteria (77 per cent). (Finding 8.1)

b) Weaknesses were found in the traceability of the non-application of the quality standards. (Finding 9)

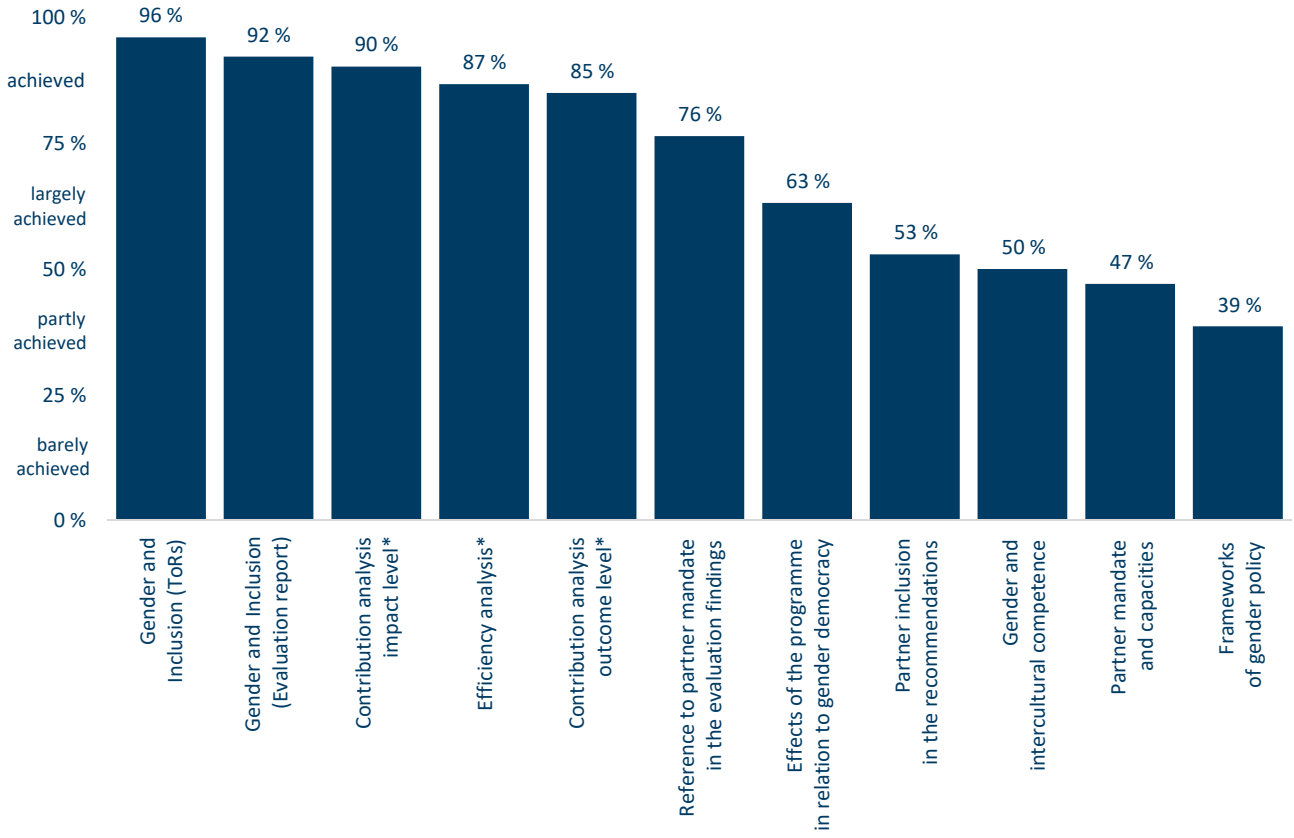
- The organisational documents did not clearly identify and prescribe all the organisation-specific quality criteria. (Finding 9.1)
- One quality criterion in the area of "gender" and one in the area of "partner role" were each "partly achieved". (Finding 9.2)
- Non-application was not documented or explained in the evaluations. (Finding 9.3)

To analyse their application, the quality criteria first of all had to be operationalised. The thresholds for all but two of the eleven organisation-specific quality criteria were at least "largely achieved". The DRK, EWDE, GIZ and hbs obtained an average score for application of around 71 per cent, which is slightly less than their average score for application of the OECD-DAC and DeGEval standards. This shows that organisation-specific quality criteria are relevant, and that they were also applied in the evaluations. The quality criteria most frequently applied were those in the area of "gender and inclusion", and the three GIZ quality criteria "methods of contribution analysis at impact level", "methods of contribution analysis at outcome level" and "efficiency analysis (follow-the-money approach)" (Figure 17). Two quality criteria were applied to a degree of less than 50 per cent in the evaluation documents of the respective organisation. It remains to be seen to what extent an average application of about 71 per cent will meet the organisations' requirements in the future, or whether it would make sense to adjust the benchmark upwards. As with the internationally recognised quality criteria, no explanations were available for non-application at the evaluation level.

On average, the GIZ[75] achieved the three organisation-specific quality criteria (approximately 87 per cent). This value was roughly in line with the average finding for the GIZ regarding the application of the OECD-DAC and DeGEval quality criteria (approximately 86 per cent). Thus, the picture was also positive for GIZ. Non-application was not specified in the evaluation documents. There is a weakness here.

75 For reasons of anonymity, only the findings for the GIZ are explained in more detail here and in the graph.

Figure 17 Achievement of the organisation-specific quality standards



Source: DEval, authors' own graph

*Note: * = quality criteria of the GIZ. ToR = terms of reference. The assignment of the quality criteria to the other non-governmental organisations is anonymised, in accordance with the other findings in section 4.2.*

4.2.4 Comparison with the sustainability meta-evaluation (GIZ and KfW)

This fourth sub-section describes the findings on application of the quality criteria of the sustainability meta-evaluation for GIZ and KfW.

This meta-evaluation builds on the quality criteria and findings generated in the sustainability meta-evaluation. The sustainability meta-evaluation by Noltze et al. (2018) was a cross-organisational meta-evaluation of GIZ and KfW evaluations. It examined the quality of the evaluation practice of 513 GIZ and KfW evaluations conducted between 2010 and 2016. Like the present meta-evaluation, it dealt with the exchange and examination of quality in evaluation practice with a focus on two official implementing organisations. The 16 quality criteria examined in the sustainability meta-evaluation focus on the area of methodology and thus on aspects of the standard cluster "reporting and methods". In the present meta-evaluation, 15 of these quality criteria were taken up and re-examined for 106 GIZ[76] and KfW evaluations carried out between 2016 and 2020. This made it possible to analyse (i) the current status of the application of these quality criteria, and (ii) any difference in application arising since the last analysis. Box 7 shows the general conclusion and the main findings of the investigations carried out (the names of the quality criteria of the sustainability meta-evaluation were retained).

76 For the GIZ, the figures refer to the years 2018 to 2020.

Box 7 General conclusion on the application of the sustainability meta-evaluation quality criteria

To what extent are strengths and weaknesses evident in the application of the sustainability meta-evaluation quality criteria in the evaluations of the GIZ and KfW? (Evaluation question 2c)

Regarding the application of these quality criteria, the picture was a positive one – with a few exceptions. Specifically, the quality criteria were achieved on average to a degree of about 75 per cent. This represents a somewhat higher degree of application than was found for the OECD-DAC and DeGEval standards. However, challenges remained in the application of the quality criteria "selection procedure for interviewees described" and "control/comparison groups included". Furthermore, it should be noted that the application of all the quality criteria has improved – in some cases clearly – since the sustainability meta-evaluation. Overall, an average difference of 36 per cent was observed. These changes indicate that the measures implemented after the sustainability meta-evaluation to improve the evaluation practices of GIZ and KfW might have affected application. This is a very positive result in light of the extensive efforts made by a large number of actors in connection with the measures. It should be noted, however, that alternative explanations could not be ruled out (for example, operationalisation of the quality criteria in ways that make it relatively easy to achieve the benchmarks, or changed documentation methods).

a) Strengths were evident in the current application of 13 quality criteria, and the positive change in application since the sustainability meta-evaluation. (Finding 10)

- On average, the quality criteria were achieved to a degree of approximately 75 per cent. Eight out of 15 quality criteria were achieved and five were largely achieved. (Finding 10.1)
- Between the sustainability meta-evaluation (t1) and the present meta-evaluation (t2), the application of the quality criteria increased by 36 per cent on average; the differences in seven quality criteria are statistically significant and meaningful. (Finding 10.2)

b) Weaknesses were found in the current application of two quality criteria. (Finding 11)

- On average, the quality criteria "selection procedure for interviewees described" and "control/ comparison groups included" were partly achieved and barely achieved respectively. (Finding 11.1)

On average, the quality criteria were achieved to a degree of about 75 per cent. Eight quality criteria were achieved in full, five were largely achieved, and one each were partly and barely achieved respectively; the latter are "selection procedure for interviewees described" and "control/comparison groups included". The average application of these quality criteria was thus slightly higher than the average application of the OECD-DAC and DeGEval standards by the GIZ and KfW. The two least applied quality criteria were achieved to a degree of less than 50 per cent. The quality criterion "control/comparison groups included", which examined whether the effects of the development intervention were achieved based on a comparison between the control group (outside the scope of influence of the development intervention) and the target group (within the scope of influence of the development intervention), was applied to a noticeably lesser extent – 18 per cent – than the quality criterion "selection procedure for interviewees described" (43 per cent).

Overall, the quality criteria were applied on average about 36 per cent more at the time of this analysis than at the time of the sustainability meta-evaluation. This showed a clear difference in the application of the methodologically oriented quality criteria by the GIZ and KfW. This difference is statistically significant and meaningful for seven quality criteria. The clearly positive difference in findings between the present meta-evaluation and the sustainability meta-evaluation indicates an improvement in the application of the quality criteria (Figure 18). This difference varies between 8 per cent for the quality criterion "object (intervention) described" and 91 per cent for "causality inferred on the basis of plausibility". The difference between the individual quality criteria is presumably related to the degree of application at the time it was first measured. (For example, the quality criterion "object [intervention] described" was already being applied at a rate of 92 per cent at the time of the sustainability meta-evaluation. This meant that a possible improvement could not have gone beyond 8 per cent.) The seven quality criteria that have improved to a statistically significant degree are "object (intervention) described", "before-and-after comparison",

"causality inferred on the basis of plausibility", "database adequate for conclusions", "Theory of Change largely operationalised through indicators", "interviewees identified" and "conclusions from data plausibly substantiated".[77] Since the quality criterion "control/comparison groups included" continued to be barely applied – the application changed from approximately 9 to 18 per cent – a systematic non-application of the quality criterion in GIZ and KfW project evaluations cannot be ruled out. Given that (i) KfW conducted ex post evaluations, (ii) there was no obligation for either organisation to integrate or consider integrating control/comparison groups in evaluations (e.g. in the form of rigorous impact evaluations), and (iii) it is for example virtually impossible to implement a control/comparison group design in policy advice interventions, this could explain the low finding for application of the quality criterion.

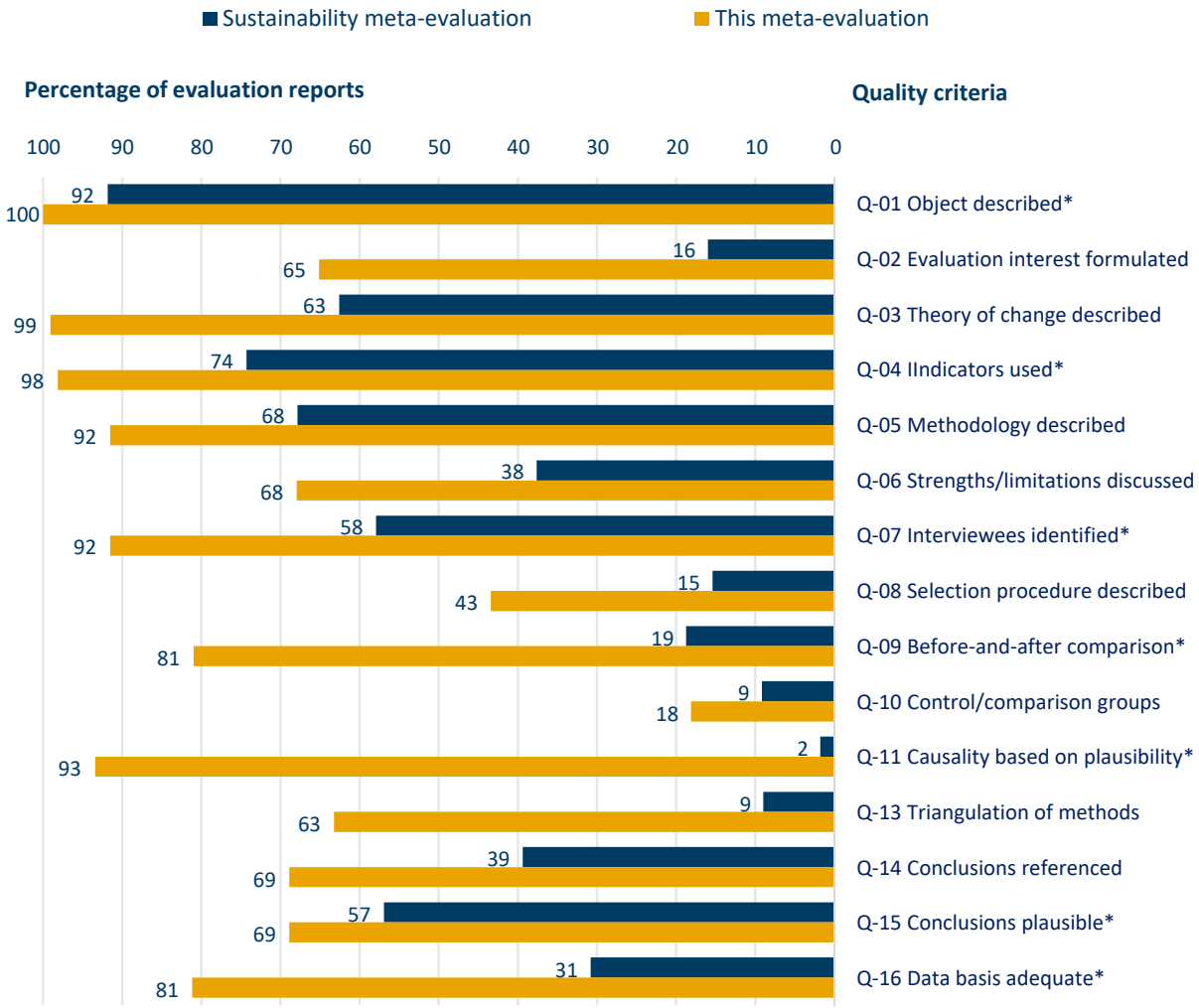
The positive difference could be related to measures to increase the quality of project evaluations by the GIZ and KfW following the sustainability meta-evaluation, but alternative explanations are also possible. Both implementing organisations had applied far-reaching measures through their own reforms of their evaluation systems, based on the recommendations from the sustainability meta-evaluation, and with support from the BMZ and DEval. GIZ had already launched a reform of its evaluation system in mid-2017 (GIZ, 2018a). It continued to do so in line with the findings and recommendations of the sustainability meta-evaluation. One result of this has been that more project evaluations, which had previously been carried out on a decentralised basis, are now being coordinated centrally by the Evaluation Unit and implemented by independent external experts (GIZ, 2018a). Furthermore, it is now standard practice in central project evaluations to conduct contribution analyses (GIZ, 2018b). This could explain the 92 percent increase in the application of the quality criterion "causality inferred on the basis of plausibility".[78] Meanwhile, KfW has revised the format of ex post evaluations along the lines of "Rapid Appraisal 2.0" (KfW, 2019)[79]. It has increased the transparency of the methodological approach, and implemented supplementary procedures for evaluating conclusions and lessons learned. These measures could explain the positive difference in the application of the quality criteria. However, it cannot be ruled out that the difference is also due to the way in which the quality criteria were operationalised in the sustainability meta-evaluation, which makes them easy to achieve in some cases. It is also possible that the GIZ and KfW were already applying the quality criteria at the time of the first meta-evaluation, but that this was not described clearly and logically. The difference might then be attributable to more systematic documentation. Finally, something that also cannot be ruled out is that the discrepancy might have arisen due to a combination of these explanations.

77 For details of the findings from the structural equation model, see section 4.1.3 of the online annex.

78 The quality criterion is: "The criterion is met when the results of the development intervention are inferred using a systematic procedure based on plausibility (especially theory-based approaches, e.g. contribution analysis)."

79 This refers in particular to DEval recommendation 1 on the further development of evaluation practice: "Given the growing demands placed on evaluation as a tool for learning and accountability, the GIZ and KfW should develop measures to ensure that exhaustive use is made of further potential to increase the quality of evaluation, particularly with respect to substantiating results and sustainability."(Noltze et al., 2018: 47).

Figure 18 Percentage of evaluation reports by quality criteria achieved at both points in time



Source: authors' own graph based on Noltze et al. (2018: 28, figure 4)

Note: Sustainability meta-evaluation = 513 evaluations, present meta-evaluation = 106 evaluations; Q-12 data triangulation was not included in the present meta-evaluation, as it was covered by the quality criterion "triangulation of methods". The names of the quality criteria were taken from the sustainability meta-evaluation. Since some of the OECD-DAC and DeGEval quality criteria in the present meta-evaluation had to be transformed in order to include them in the comparison with the quality criteria of the sustainability meta-evaluation, the findings are not always congruent with those in section 4.2.1. These quality criteria have improved statistically significantly ($p < 0.05$) and meaningfully. For further details see section 4.1.3 of the online annex.*

4.3 Explaining the application of the quality criteria

This section presents the extent to which the factors identified in the literature and the focus group discussions explain or are linked to the application of the quality criteria. Box 8 shows the general conclusion and the main findings of the analyses.

Box 8 General conclusion explaining the application of the quality standards

To what extent are country-specific, evaluation-specific and organisation-specific factors linked to the application of quality standards? (Evaluation question 3)

a) Overall, the findings gave little indication of meaningful links between the identified factors and the quality criteria. Some meaningful positive links were identified in the evaluation-specific dimension between the factors "number of internal and external evaluators", "involvement of internal and external stakeholders" and "year of evaluation", and various quality criteria. Since many models were assessed in the analyses, some with intercorrelating quality criteria, the findings must be interpreted with caution.

b) Few significant findings were obtained in the regression analyses.[80] (Finding: 12)

- In the evaluation-specific dimension, a positive correlation with the application of various quality criteria was found both with the "competence of the evaluators" (especially the proxy "number of internal and external evaluators") and with "quality assurance processes" (especially the proxy "stakeholder involvement"). Furthermore, the "year of evaluation" also displayed positive correlations with two quality criteria. (Finding: 12.1)
- Country-specific factors were not significantly linked to the application of the quality criteria. (Finding: 12.2)
- The organisation-specific factors did not show a clear picture regarding links to the application of the quality criteria. (Finding: 12.3)

The information below sheds light on the identified factors or their proxies, in order to contextualise the findings of the analysis. Approximately 81.1 per cent of the evaluations were conducted exclusively by external evaluators, and 13.2 per cent exclusively by internal evaluators (persons working within the respective organisation). In 5.6 per cent of cases, there were one or more internal evaluators in addition to external evaluators. Overall, the evaluations were conducted by at least one and up to ten evaluators (average: 1.9 evaluators). The average daily rate of the external evaluators was about 440 euros, the average ratio between evaluator days and the financial volume of the development intervention was about 340,000 euros/day. Remote data collection took place in 10 per cent of the evaluations, semi-remote data collection in 6 per cent, and on-site data collection in 84 per cent of the evaluations.[81] Table 6 shows an overview of the findings from the regression analyses.[82]

Some factors could only be examined empirically through proxy variables. This was because either there was no uniform cross-organisational definition or no causally attributable links, or because data were unavailable. These factors included "cultural context", "pandemic", "competence of the evaluators", "evaluation costs", "quality assurance processes", "structured planning process", "evaluation and learning culture" and "evaluation unit in place".

The findings from the regression analyses showed no significant empirical links between the country context and the examined quality criteria or the standard cluster "reporting and methods". Based on the selected model specifications, only the "social capital index" showed a negative correlation with the

80 For further details on the regression analyses, see section 4.2 of the online annex.

81 Descriptive information on the organisational context is presented in section 1.1.

82 For details on the regression analyses, see section 4.2 of the online annex.

"description of the methodological adequacy", though this was not negatively meaningful. A possible explanation for the lack of any correlation with "pandemic" could be that the time period of the meta-evaluation only includes the beginning of the pandemic (the year 2020), and thus potential effects on the quality of evaluations could not yet be captured. The regression analyses confirmed the findings of Wencker and Verspohl (2019) that "conflicts" are not linked to quality criteria in the standard cluster "reporting and methods". The findings also showed that positive – though not significant – correlations exist with individual quality criteria in the standard cluster "participation, independence and fairness".

In the evaluation dimension, the factors "competence of the evaluators" and "quality assurance processes" were positively linked to individual quality criteria. In addition, the "year of evaluation" was positively linked to some quality criteria in the standard cluster "reporting and methods". The "competence of the evaluators", operationalised among other things through the "number of internal and external evaluators", was positively and in part meaningfully linked to selected quality criteria in the standard cluster "reporting and methods" ("description of the methodological adequacy" and "information content of the executive summary"). Aspects of the "quality assurance processes" of an evaluation (for example "stakeholder involvement") were also positively and meaningfully linked to individual quality criteria ("consideration of the context", "clarity of the information sources" and "summative index for reporting and methods"). The "year of evaluation" was positively – and in some cases meaningfully – linked to several quality criteria of the standard cluster "reporting and methods", as well as the summative and the factor index of the standard cluster, while the "clarity of the information sources" was negatively linked. Similar to a meta-evaluation of decentralised evaluations from 2017 to 2020 from Finland by Väth et al. (2022), the "information content of the terms of reference" was also found to be a significant factor, but in contrast to the Finnish meta-evaluation not a strong one. Other significant but non-meaningful[83] findings are: The "information content of the terms of reference" was also negatively linked to "information content of the executive summary". The proxy for the evaluation costs (evaluator days in relation to the financial volume of the development intervention) was negatively linked to the quality criterion "coherence of data-findings-conclusions". While unlike in the case of Hundt and Bräuer (2021) no relationship was found between "remote data collection" and the "description of the methodological adequacy", a negative correlation was found between "remote data collection" and "evaluation capacity development in the partner country".

The organisation-specific factors did not show a clear picture regarding links to the application of the quality criteria. The findings on the factors in the organisational context did not show any meaningful links to individual quality criteria. There was a positive – non-meaningful – correlation between "evaluation activity" and a more informative "summary", and a negative one to "description of the methodological adequacy". Furthermore, there was a negative link between the "size of evaluation units/desks" and the "information content of the terms of reference". The negative links were only present for individual quality criteria. In addition, some of the organisations had statistically significant links to the quality criteria.

For the quality criteria "description of the methodological adequacy" and "description of the Theory of Change", both of which were only "partly applied" on average (see section 4.2), significant, albeit non-meaningful, links were found. For the quality criterion "description of the Theory of Change", a positive link to the "year of evaluation" was found. In other words, over the years 2016 to 2020, the cause-effect relationships were better presented in the organisations' evaluations. Furthermore, a higher "number of evaluators" and the "year of evaluation" are both correlated with a better "description of the methodological adequacy". The link is strongest when four evaluators carried out the evaluation as opposed to one. There are also negative links to the "social capital index" and the "evaluation activity".

.

[83] The links between quality criteria and factors that had a sufficiently large effect size were considered meaningful. The effect sizes were determined for the (ordered) logistic models (see Table 6 and section 4.2 of the online annex).

Table 6 Links between the factors and the application of the quality criteria

		SC B & M												SC N	SC P, U & F		
Factor	Proxy	Evaluation object	Context*	ToC	Focus of evaluation	Information sources	Description of methodology	Coherence	ToRs	Summary	Existence of IR	Summative index SC R & M	Factor index SC R&M	Recommen-dations*	Evaluation capacity development	Partner-country Evaluators	Stakeholder involvement
	Regression-model no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Fragility	Conflict	–	–	–†	–	–	–	–	–	–	–	–	–		–	–	–
Cultural context	Social capital	+	+	+	+	+	–*	+	+	+	+	–†	–†				+
Pandemic	Year 2020	–	–	–	–	–	–	–	–	–	–	–	–		–	–	–
Competence evaluators	No. of evaluators	+	+	+	+	+	+*	+	+	+***k	+	+†	+		+	+	
	Daily rate evaluators	+	+	+	+	+	+	+	+	+	+	+	+		+		+
Quality assurance	ToRs	+	+	+	+	+	+	+*		–*	+			+			
	Existence of IR	+	+	+	+	+†k	+k	+	+	+			+	+*	–*		+
	Stakeholder involvement	+	+**	+†	+	+***k	+	+	+†k	+	+	+*	+				
Remote data collection							–								–*	+	
Costs	Evaluator days/costs for intervention	+	+	+	+	+	+	–**	+	+	+	+	+				
Year of evaluation		+	+*	+***	+**k	–*	+*	+	+**k	+	+	+*	+*				
Evaluation units/desks absolute		+	+	+	+	+	–	+	–*	+	+	+	+	+	+	+	+

		SC B & M												SC N	SC P, U & F		
Factor	Proxy	Evaluation object	Context*	ToC	Focus of evaluation	Information sources	Description of methodology	Coherence	ToRs	Summary	Existence of IR	Summative index SC R & M	Factor index SC R&M	Recommen-dations*	Evaluation capacity development	Partner-country Evaluators	Stakeholder involvement
	Regression model no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Evaluation units/desks relative to evaluations		+	+	+	+	+	+	+	+	+	+	–†	+	+	+	+	+
Evaluation units/desks relative to organisation		+	+	+	+	+	+	+	+†m	+	+	+	+	+	+	+	+
Evaluation activity		+	+	+	+	+	–*	+	+	+*	+	+	+	+	+	+	+
Evaluation and learning culture	Organisation[1]	+	+	+	+†	+	+*	+*	+**	+	+	+*	+*	+	+†	+†	+*

Source: DEval, authors' own table

*Note: SCs = standard clusters; R & M = reporting and methods; U = usability; P, I & F = participation, independence and fairness; blue = positive link; orange = negative link; † = marginal correlation; grey = no statistically significant link; blank = not tested. † < 0.10; * = p < 0.05; ** = p < 0.01; *** = p < 0.001; k = small (odds ratio between 1.68 and 3.45) and m = medium (odds ratio between 3.46 and 6.70) effect size (Chen and Cohen, 2010). The findings for the control variables are not shown. [1]For reasons of anonymity, the links between the individual organisations and the quality criteria are not shown. Instead, the statistical links show the average for the findings. *For "context" and "usefulness of recommendations", the two quality criteria were combined into one quality standard (as a summative index) in each case.*

5. CONCLUSIONS AND RECOMMENDATIONS

This meta-evaluation provides cross-organisational findings on the understanding of quality, and the application of selected quality standards, in evaluations of official implementing (governmental) and non-governmental German development cooperation organisations. It also supplies findings on possible factors linked to the application of quality criteria. *This will enable (cross-cutting) learning. Given the cross-organisational nature of the meta-evaluation, subsequent studies (such as internal meta-evaluations by organisations) will be able to use it. They can draw from it and consider findings not only on the degree of application of the quality standards, but also on a wide variety of forms of, and explanations for, the application or non-application of particular quality standards. It offers all involved organisations guidance for further developing their own evaluation practice. For the official implementing organisations, it also provides an opportunity to account for their application of the quality standards externally. Finally, the OECD-DAC and DeGEval standards, which were systematically used for the analysis grid of the meta-evaluation, also represent key quality standards for the BMZ Evaluation Policy that has since come into force. The findings and the analysis grid of the meta-evaluation thus form a basis for an analysis grid that is potentially still to be developed in relation to the BMZ Evaluation Policy.*

Below is an introductory section to classify the conclusions and recommendations. Subsequently, the recommendations are divided into five thematic areas: 1) identification of relevant quality standards and systematic prescription in organisational documents, 2) identification of non-relevant quality standards and systematic prescription in organisational documents, 3) assurance and traceability of the application/non-application of relevant quality standards at evaluation level, 4) joint learning, and 5) assurance and traceability of the application/non-application of the sustainability meta-evaluation quality criteria. As both the identification and the assurance and traceability of the application of required quality standards are fundamental to ensuring good evaluations, there is a need for action by all organisations. This varies between organisations, depending on their degree of previous engagement and application/achievement. It is important to bear in mind that all quality standards of the OECD-DAC and DeGEval are to be regarded as equivalent, and should be included.

The recommendations of the meta-evaluation are derived from the findings presented in the previous chapter, and are addressed to the BMZ, as well as the BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR and PTB. Furthermore, the recommendations may also be appropriate and useful for VENRO and other non-governmental organisations. The recommendations are mainly derived from the evaluation questions on the understanding of quality and the application of selected quality standards and quality criteria (OECD-DAC, DeGEval and/or organisation-specific quality standards, as well as quality criteria of the sustainability meta-evaluation, evaluation questions 1 and 2a to 2c). The answers to evaluation question 3 (factors linked to the application of quality standards) are used as background information or as part of the implementation guidance. The recommendations are worded in general terms. This means that each involved organisation must consider its own organisation-specific findings in order to determine the particular relevance of the recommendations to it. This is because the cross-organisational mean value presented in the findings is not sufficiently meaningful for evaluating the application of quality standards by specific organisations. There are organisations, for example, that only barely or partly applied quality standards, even though the average rating for application was largely achieved or achieved.

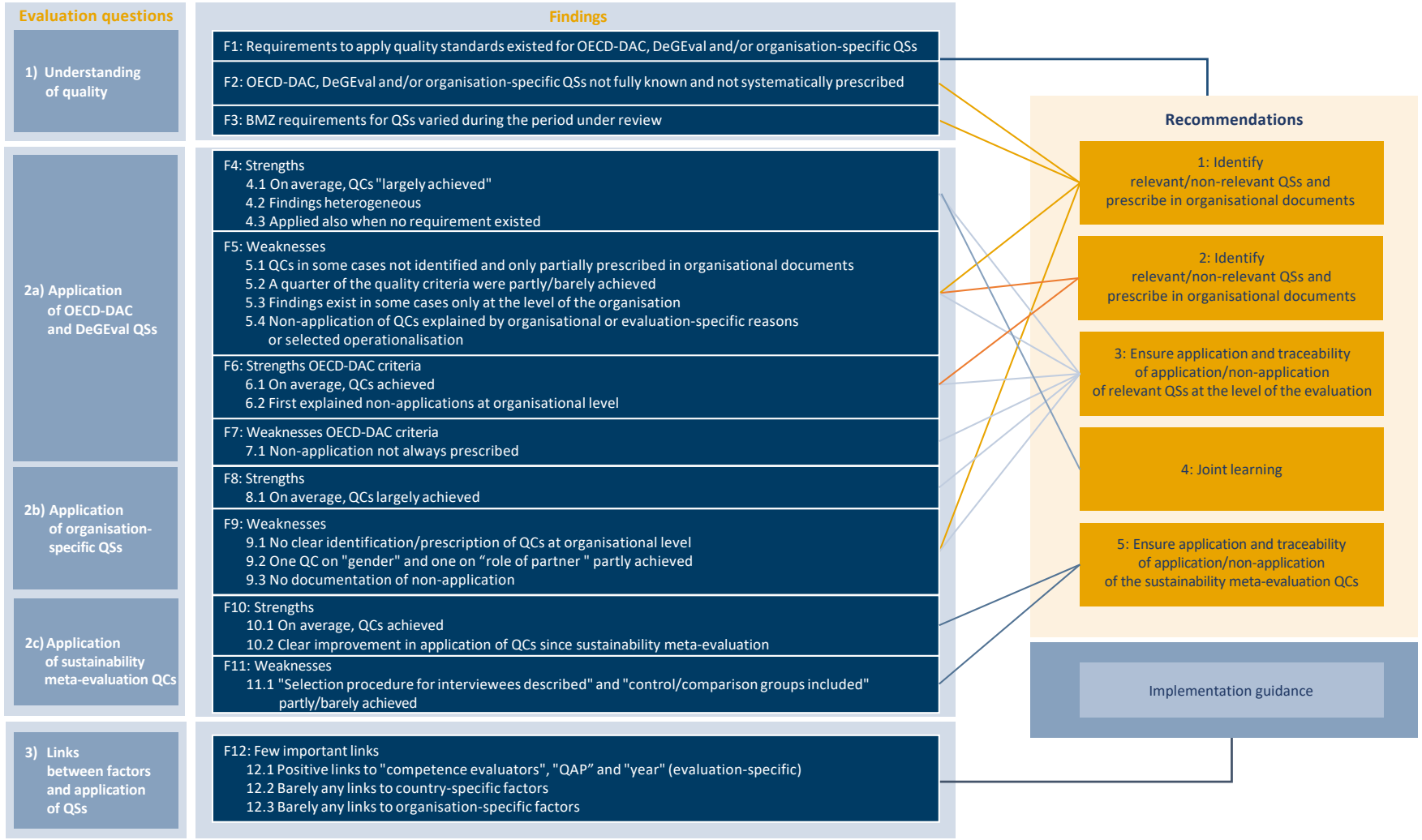
The findings for the four governmental organisations BGR, GIZ, KfW and PTB are presented separately in the annex to the report (section 7.1), so that they can be used for purposes of accountability, and planning and monitoring implementation of the recommendations of this meta-evaluation. The organisation-specific findings for the non-governmental organisations were shared exclusively with the respective responsible officers of the evaluation units/desks. Based on their organisation-specific results, the organisations can therefore identify the recommendations and implementation guidance that apply

to them.[84] The criteria-based selection of non-governmental organisations focuses on their structural heterogeneity, and thus reflects the range of possible degrees and forms of application for different organisations. This means that non-governmental organisations which were not involved can also determine the relevance of the findings to them, and thus draw guidance from the conclusions and recommendations. The recommendations addressed to the BMZ are intended for the BMZ Evaluation Division.

Figure 19 provides an overview of which findings were used to develop which recommendations. For the sake of clarity, both the findings and the recommendations are presented in abridged form.

84 The fact that some of the findings for the organisations are anonymised does not affect the need to implement the recommendations presented. This is also the case because almost all of the organisations shown anonymously are required to apply quality standards. Furthermore, the BMZ Evaluation Policy that came into force in 2021 also provides guidance for the non-governmental organisations.

Figure 19 Overview of the derivation of the recommendations from the findings



Source: DEval, authors' own graphic

Note: F = finding; QS = quality standard; QC = quality criterion; QAP = quality assurance process

Identification of relevant/non-relevant quality standards and prescription thereof in organisational documents

To ensure good evaluation practice in the long term, it is necessary to identify relevant and non-relevant quality standards for an organisation's evaluations and to systematically prescribe them in organisational documents. German development cooperation organisations are required to implement various quality standards based on a number of sources – for example, through memberships, (funding) guidelines and/or organisational documents. In addition to the OECD-DAC and DeGEval quality standards examined in this meta-evaluation, other quality standards may also be relevant to them (e.g. quality standards from the new BMZ Evaluation Policy, other standards documents that are relevant in partner countries and/or organisation-specific quality standards). In the meta-evaluation, the quality standards (not standards documents) that are required for an organisation were not fully identified and prescribed in the organisational documents by all organisations. This represents a weakness. Systematically identifying and prescribing all required quality standards in the organisation's documents is expedient in order to 1) ensure that all quality standards relevant to an organisation are identified; 2) prevent the inadvertent non-application of individual quality standards; 3) prescribe the intentional non-application of individual quality standards; 4) establish processes for the application of all quality standards relevant to the organisation in its evaluation practice, and improve them on this basis; and 5) make transparent the relevance of required internationally recognised and organisation-specific quality standards.

Organisations sometimes change their thematic and strategic focus, and adjust their evaluation activities by region or sector. Furthermore, standards documents are revised at episodic intervals (for example, the DeGEval standards), or new ones are developed. Consequently, identifying and prescribing relevant quality standards in the organisations is an ongoing task rather than a one-off matter. Or as the DeGEval standards document (DeGEval, 2016: 13) states: "Ultimately, it is an organisation's own basic attitude to evaluation that requires it to subject even tried and tested practices to continuous critical examination, and discuss these with regard to possible improvements."

Note: The recommendations may also be appropriate and useful for organisations that were not involved.

Recommendation 1

a) As part of a revision of their evaluation practice, the evaluation units/desks of the BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR and PTB should (if they have not already done so) identify the quality standards that are required for their organisation. They should explicitly prescribe these in organisational documents, and define their application in evaluation processes. The organisations should review at regular intervals the identification and systematic prescription of quality standards. When doing so, they should specify the degree of application they require for each of the quality standards. (Findings: 2, 5.1, 9.1)

b) In the context of upcoming updates of funding guidelines or ancillary provisions for individual budget items, the BMZ should make a contribution towards strengthening its Evaluation Policy as a reference document for evaluations. As part of these updates, together with the non-governmental organisations concerned the BMZ should establish and prescribe special conditions for particular organisations (e.g. as in the case of the funding guidelines for political foundations). The principle of maximum standards should be retained here. (Finding: 3)

c) Based on its Evaluation Policy, and in dialogue with the official implementing and non-governmental organisations, the BMZ should develop an analysis grid for the application of the quality standards, taking into account the analysis grid of the present meta-evaluation. It should also make this available to the official implementing and non-governmental organisations.

Recommendation 2

a) The evaluation units/desks of the BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR and PTB should explain and prescribe in organisational documents any general non-application of particular quality standards that are required for them. (Findings: 5.1, 6.2)

b) The BMZ should reach an agreement with the official implementing organisations on the application and (explained) non-application of the quality standards described in the BMZ Evaluation Policy. It should do so either in order to jointly determine non-application at the organisational level or to document discrepancies.

Implementation guidance

- To identify relevant quality standards, all quality standards documents applicable to the organisation (e.g. the BMZ Evaluation Policy that came into force in 2021, the standards documents of the OECD-DAC and/or DeGEval), as well as relevant internal organisational documents, should be consulted. It should be taken into account that different quality standards may become relevant for different evaluation types (e.g. for decentralised evaluations).
- To determine the desired degree of achievement of the relevant quality standards (thresholds), the organisation-specific findings or the threshold of this meta-evaluation can be used. Here too, the different evaluation types should be considered.
- When developing the analysis grid, the BMZ should take into account different but equally valid ways of applying quality standards, in order to do justice to the heterogeneity of the organisations.

Ensuring the application and traceability of the application/non-application of relevant quality standards at the level of the evaluation

Application of the OECD-DAC and DeGEval standards is largely institutionalised in the involved German development cooperation organisations, although the degree of application of the quality criteria varied significantly within and between the organisations. Overall, on average the OECD-DAC and DeGEval quality standards were largely achieved, the OECD-DAC criteria were achieved, the organisation-specific quality standards were largely achieved and the quality criteria of the sustainability meta-evaluation were achieved. This shows that application of the quality standards has found its way into evaluation practice. This represented a strength in application by the involved organisations.

The traceability of application and especially of non-application at the level of the evaluation needs to be improved. An (explained) non-application occurred only in rare cases. This represented a weakness of the involved organisations with regard to the application of the quality standards. If the application or non-application of individual quality criteria cannot be traced at the level of the individual evaluation, any analysis and assessment of good evaluation practice is flawed, as it is then barely possible to distinguish between a quality criterion that was not applied, and one that was applied but not documented, or one that was not applied but whose non-application was explained. Furthermore, meta-evaluations with better traceability would be easier to conduct. Since the internationally recognised standards represent maximum standards, a non-application of selected standards is already inherent. This represents a recognised approach in engagement with quality standards that should be made an established practice. According to the OECD-DAC (OECD-DAC, 2010: 5), the quality standards should be "applied sensibly and adapted to local and national contexts and the objectives of each evaluation".

Recommendation 3

a) The evaluation units/desks of the BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR and PTB should, if they have not already done so, further improve the application of the quality standards required at the organisational level (recommendation 1) in individual evaluations, and especially those quality standards that are barely or party applied. Furthermore, the application or (explained) non-application of all quality standards should be traceable at the level of each evaluation and regularly reviewed by the organisations. (Findings: 4.1, 4.2, 5.2, 5.3, 6.1, 7.1, 8.1, 9.2, 9.3)

b) The BMZ should urge the official implementing organisations to ensure the application of the relevant quality standards, and the traceability of their application/non-application, at evaluation level.

Implementation guidance

- Ensuring the application of the quality standards in each evaluation:
  1) As there is evidence that the implementation of quality assurance processes (e.g. the "involvement of internal and external stakeholders") is positively linked to the application of selected quality standards, it should be ensured that the responsible officers of the evaluation units/desks are adequately trained, and have sufficient resources to adequately implement these quality assurance processes. (Finding: 12.1)
  2) As there are indications that the "competence of the evaluators" – especially the "number of evaluators" – is linked to a better application of selected methodological quality standards, organisations should consider engaging teams of evaluators for evaluations, where possible. (Finding: 12.1)
  3) Various good practice examples of the involved organisations regarding their application of selected quality standards can be found both in the report, and in the online annex (section 4.1.1). These can be used as templates for organisation-specific actions.
- Traceability of the application/non-application of the quality standards in each evaluation:
  1) Since individual quality standards are documented at different points in the evaluation process, and by different persons, the processes for application/non-application of the individual quality standards should be formulated at evaluation level.
  2) Traceability should be as lean and efficient as possible. This could be achieved, for example, by making application/non-application traceable in the organisational processes implemented for each evaluation, or through documentation at a higher level in the organisation (for example in monitoring).[85]

Joint learning

The structural heterogeneity of the involved organisations is reflected in the different understandings of quality and degrees of application of individual quality standards. Underlying this are different practices and different experiences with identifying, prescribing, assuring and tracing the application/non-application of quality standards. These offer the organisations an opportunity to learn

[85] The findings, conclusions and recommendations for the GIZ and KfW in relation to evaluation question 2c (To what extent are strengths and weaknesses evident in the application of the quality criteria of the sustainability meta-evaluation in the evaluations of the GIZ and KfW?) are presented in section 4.2.4 and chapter 5, and are not discussed again here. Evaluation question 3 (To what extent are country-specific, evaluation-specific and organisation-specific factors linked to the application of quality standards?) was not examined separately for the official implementing organisations, as no divergent causal relationships between the examined explanatory factors and quality standards were assumed for these organisations. Accordingly, no further findings are included here.

from each other. Since the organisations have gained a wide range of experience in identifying, prescribing, assuring and tracing the application/non-application of the quality standards, a cross-organisational dialogue (e.g. sharing of good practice examples) can enable institutional learning, and thus an improvement in the application of the quality standards. The involved organisations have already stated that the sharing of lessons learned to date is a good learning opportunity that should be utilised. A systematic dialogue between the organisations and the BMZ to promote a common awareness of standards also makes sense against the background of a BMZ guideline analysis grid still to be developed.

Recommendation 4

a) The evaluation units/desks of the BGR, CARE, DRK, DVV, EWDE, GIZ, hbs, KAS, KfW, MISEREOR and PTB, and representatives of VENRO, should regularly share their various lessons learned in identifying, prescribing, assuring and tracing the application/non-application of all quality standards. This dialogue should also integrate non-involved organisations and include further types of evaluation – such as decentralised evaluations – in order to continue improving the application of quality standards. (Findings: 4.2, 4.3)

b) The BMZ should financially support the dialogue with and between the organisations on identifying, prescribing, assuring and tracing the application/non-application of the quality standards.

Implementation guidance

- A regular dialogue could take place, for example, in conjunction with the annual meeting of the evaluation units/desks, in selected networks, in working groups or at forums. The latter can also be identified outside of the development cooperation context. Here, it may make sense to conduct the dialogue on a smaller scale, for example only with organisations falling under one budget item.
- The financial support relates in particular to organisations under the budget item "private German organisations" and VENRO.

Ensuring the application and traceability of the application/non-application of the quality criteria from the sustainability meta-evaluation

A positive picture emerged for the current achievement of the quality criteria from the sustainability meta-evaluation. Moreover, the application of these quality criteria has increased considerably in GIZ and KfW evaluations since the sustainability meta-evaluation. On average, current application of the quality criteria is achieved. The improvement in the application of the sustainability meta-evaluation quality criteria can presumably be attributed inter alia to its findings and recommendations, as well as to the reforms of GIZ's and KfW's evaluation practices implemented with the support of BMZ and DEval. However, other explanations are also possible, such as relatively easy-to-achieve ways of operationalising the quality criteria, or a change in the way documentation is carried out. The quality criteria examined in the sustainability meta-evaluation focus on the methodological approach of GIZ and KfW evaluations. Since methodology is also a focus of the BMZ Evaluation Policy, it makes sense to compare the quality standards of the two documents, and possibly adopt the quality criteria for the BMZ analysis grid.

The recommendations may also be appropriate and useful for organisations that were not involved.

Recommendation 5

a) When developing the analysis grid for the quality standards described in the BMZ Evaluation Policy (recommendation 1), the BMZ should consider adopting the quality criteria from the sustainability meta-evaluation. If appropriate, it should also include them in the analysis grid.

b) Based on recommendation 5a, the GIZ and KfW should ensure/improve the application/non-application of the quality criteria from the sustainability meta-evaluation that have been incorporated into a BMZ analysis grid. They should also ensure the traceability of the (explained) application/non-application for each evaluation. (Findings: 10.1, 10.2, 11.1)

6. REFERENCES

AfrEA (2020), *The African Evaluation Guidelines 2020 Version*, African Evaluation Association, Washington, D. C.

Backhaus, K. et al. (2011), *Multivariate Analysemethoden: eine anwendungsorientierte Einführung*, Springer, Berlin, 13., revised edition.

Backhaus, K. et al. (2015), *Fortgeschrittene multivariate Analysemethoden: eine anwendungsorientierte Einführung*, Springer Gabler, Berlin, 3. edition.

Beywl, W. und M. Niestroj (2009), *Das ABC der wirkungsorientierten Evaluation: Glossar - deutsch/englisch - der wirkungsorientierten Evaluation*, Univation - Institut für Evaluation Dr. Beywl und Associates, Köln, 2. edition.

BMF (2020), *Bundeshaushaltsplan 2020. Einzelplan 23. Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung*, Federal Ministry of Finance, Berlin/Bonn.

BMZ (2006), *Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit. Eine Orientierung für Evaluierungen des BMZ und der Durchführungsorganisationen*, Federal Ministry for Economic Cooperation and Development, Bonn/Berlin.

BMZ (2016), *Richtlinien zu Förderung entwicklungswichtiger Vorhaben der politischen Stiftungen aus Kapitel 2303 Titel 68704*, Federal Ministry for Economic Cooperation and Development, Bonn/Berlin.

BMZ (2020), *Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit. BMZ-Orientierungslinie zum Umgang mit den OECD-DAC-Evaluierungskriterien in Evaluierungen der deutschen bilateralen Entwicklungszusammenarbeit*, Federal Ministry for Economic Cooperation and Development, Bonn/Berlin.

BMZ (2021a), *Evaluierung der Entwicklungszusammenarbeit: Leitlinien des BMZ*, Federal Ministry for Economic Cooperation and Development, Bonn/Berlin.

BMZ (2021b), "Leitlinien für die bilaterale finanzielle und technische Zusammenarbeit mit Kooperationspartnern der deutschen Entwicklungszusammenarbeit", BMZ Konzepte, Nr. 165, Federal Ministry for Economic Cooperation and Development, Bonn/Berlin.

Borrmann, A. et al. (1999), *Erfolgskontrolle in der deutschen Entwicklungszusammenarbeit: Analyse, Bewertung, Reformen*, Nomos, Baden-Baden.

Borrmann, A. und R. Stockmann (2009), *Evaluation in der deutschen Entwicklungszusammenarbeit. Band 1: Systemanalyse*, Waxmann, Münster.

Brown, T. A. (2006), *Confirmatory Factor Analysis for Applied Research*, Guilford Press, New York.

Caracelli, V. J. und L. J. Cooksy (2009), „Metaevaluation in Practice", *Journal of MultiDisciplinary Evaluation*, Vol. 6, Nr. 11, S. 1–15.

Caspari, A. (2010), *Lernen aus Evaluierungen. Meta-Evaluation & Evaluationssynthese von InWEnt-Abschlussevaluierungen 2009*, Capacity Building International, Bonn.

Caspari, A. (2011), *Meta-Evaluation & Evaluationssynthese 2011 - Hauptbericht*, o.V. Frankfurt am Main.

Caspari, A. (2012), *Meta-Evaluation, Evaluationssynthese, Evaluation Review and Systematic Review – eine Begriffsklärung,* Frankfurt University of Applied Sciences, Frankfurt am Main.

Chen, H. et al. (2010), "How Big Is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies", *Communications in Statistics - Simulation and Computation,* Vol. 39, No. 4, pp. 860–864.

Church, C. und J. Shouldice (2002), *The Evaluation of Conflict Resolution Interventions: Framing the State of Play*, International Conflict Research, Derry/Londonderry.

DeGEval (2016), "Standards für Evaluation", Evaluation Society, Mainz.

DeGEval (2021), "DeGEval Beitrittsantrag", Evaluation Society, Mainz.

DEval (2020), *Evaluierungskriterien für Evaluierungen des Deutschen Evaluierungsinstituts der Entwicklungszusammenarbeit*, German Institute for Development Evaluation, Bonn.

DEval (2021a), *Ablauf einer DEval-Evaluierung - Zur Rolle der Referenzgruppe*, German Institute for Development Evaluation, Bonn.

DEval (2021b), *DEval-Evaluierungen 2021 - 2023*, German Institute for Development Evaluation, Bonn.

Döring, N. und J. Bortz (2016), *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*, Springer, Berlin, 5. edition.

FES (2015), *Metaevaluierung: Evaluierungen in der Internationalen Entwicklungszusammenarbeit der Friedrich-Ebert-Stiftung*, Friedrich-Ebert-Stiftung, Berlin.

Freimann, I. und M. Krämer (2016), *Meta-Evaluierung der Projektevaluierungen (PEV)*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn/Eschborn.

Freimann, I. und M. Krämer (2017), *Querschnittsauswertung (QSA) von Projektevaluierungen (PEV) 2016 Meta-Evaluierung*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn/Eschborn.

GIZ (2018a), *Kontrolle ist gut: Höhere Standards für Evaluierungen*, Deutsche Gesellschaft für Internationale Zusammenarbeit, https://www.giz.de/de/mediathek/66304.html (accessed on 19.06.2022).

GIZ (2018b), *Das Evaluierungssystem der GIZ: Zentrale Projektevaluierungen im BMZ-Geschäft*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn/Eschborn.

von Gumppenberg, M.-C. et al. (2022), „Remote-Evaluation: Erfahrungen bei der Umsetzung von Remote-Evaluationen im Bereich der Entwicklungszusammenarbeit und Humanitären Hilfe – Genese, Stärken, Schwächen und Ausblick", *Zeitschrift für Evaluation*, Nr. 1/22.

Hageboeck, M. et al. (2013), *Meta-evaluation of quality and coverage of USAID evaluations 2009-2012*, United States Agency for International Development, Washington, D. C.

HTSPE LTD. (2011), *Mid-term Meta Evaluation of IPA Assistance Evaluation Report*, EU Commission, Brussels.

Hundt, V. und B. Bräuer (2021), *Remote und Semi-Remote – Erfahrungen bei der Durchführung zentraler Projektevaluierungen*, Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn/-Eschborn.

KfW (2019), *„15. Evaluierungsbericht 2017–2018"*, KfW Development Bank, Frankfurt am Main.

Koy, J. et al. (2016), *Meta-Evaluierung der Projektevaluierungen aus den Jahren 2014-2015*, Misereor, Aachen.

Krämer, M. und O. Almqvist (2019), *Meta-Evaluierung und statistische Auswertung der Projektevaluierungen 2017 / 2018 - Teil II Statistische Auswertung*, Bonn.

Krippendorff, K. (2012), *Content Analysis: An Introduction to its Methodology*, SAGE, Thousand Oaks, CA, 2. edition.

Kuckartz, U. (2014), *Mixed Methods: Methodologie, Forschungsdesigns und Analyseverfahren*, Springer, Wiesbaden.

Lange, S. et al. (2020), *Remote evaluation. Initial Experience and Recommendations*, National Metrology Institute, Braunschweig.

Lücking, K. et al. (2015), *Evaluierungspraxis in der deutschen Entwicklungszusammenarbeit. Umsetzungsmonitoring der letzten Systemprüfung und Charakterisierung wesentlicher Elemente*, German Institute for Development Evaluation (DEval), Bonn.

Mäder, S. (2020), *Methoden als situierte Praxis: Die Gruppendiskussion in der Programmevaluation*, University of Hildesheim, Hildesheim.

Mauthofer, T. und S. Silvestrini (2018), *Meta-Evaluation of 33 Evaluation Reports of World Vision Germany*, World Vision Germany, Saarbrücken.

Morgan, D. L. (1999), *The Focus Group Guidebook*, SAGE, Thousand Oaks, CA.

Noltze, M. et al. (2018), *Meta-Evaluierung von Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit*, German Institute for Development Evaluation (DEval), Bonn.

OECD (2012), *„Evaluating Peacebuilding Activities in Settings of Conflict and Fragility - Improving Learning for Results“*, DAC Guidelines and Reference Series, Nr. 40, Organisation for Economic Co-operation and Development, Paris.

OECD (2013), *The DAC Network on Development Evaluation – 30 Years of Strengthening Learning in Development,* Organisation for Economic Co-operation and Development, Paris.

OECD (2021), *DAC-Prüfbericht über die Entwicklungszusammenarbeit: Deutschland 2021 (Kurzfassung): Wichtigste Ergebnisse und Empfehlungen*, Organisation for Economic Co-operation and Development, Paris.

OECD (2022), *Recommendation of the Council on OECD Legal Instruments Public Policy Evaluation*, Organisation for Economic Co-operation and Development, Paris.

OECD-DAC (2002), *Glossary of Key Terms in Evaluation and Results Based Management*, Organisation for Economic Co-operation and Development, Development Assistance Committee, Paris.

OECD-DAC (2010), *Qualitätsstandards für die Entwicklungsevaluierung*, Organisation for Economic Co-operation and Development, Development Assistance Committee, Paris.

Queiroz de Souza, A. (2017), *Meta-Evaluation and Analysis of Project Evaluations 2016*, Welthungerhilfe, Bielefeld.

Rodríguez Bilella, P. et al. (2016), *Evaluation Standards for Latin America and the Caribbean*, Renewable Energy for Latin America and the Caribbean, Fomento de Capacidades en Evaluación, Buenos Aires.

Seawright, J. und J. Gerring (2008), „Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options“, *Political Research Quarterly*, Vol. 61, Nr. 2, S. 294–308.

Silvestrini, S. und S. Bäthge (2019), *Meta-Evaluation of ADA Project and Programme Evaluations - Executive Summary*, Austrian Development Agency, Saarbrücken/Wien.

Silvestrini, S. et al. (2018), *Meta-evaluation of Project and Programme Evaluations in 2015–2017*, Ministry for Foreign Affairs of Finland, Saarbrücken/Helsinki.

UNDAF (2017), *UNDAC Companion Guidance: Theory of Change*, United Nations Development Group, New York.

UNEG (2016), *Norms and Standards for Evaluation*, United Nations Evaluation Group, New York.

UNFPA (2020), *UNFPA Evaluation Office - Assessing the Quality of Developmental Evaluations at UNFPA*, United Nations Population Fund, New York.

Väth, S. J. et al. (2022), *Evaluation: Metaevaluation of MFA's Project and Programme Evaluations in 2017-2020*, Ministry for Foreign Affairs of Finland, Saarbrücken/Helsinki.

Weiber, R. und D. Mühlhaus (2010), *Strukturgleichungsmodellierung: eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS*, Springer, Berlin.

Wencker, T. und I. Verspohl (2019), *German Development Cooperation in Fragile Contexts*, German Institute for Development Evaluation (DEval), Bonn.

World Bank (2020a), *Summary of Lessons from an Informal Conversation: Challenges of Conducting Remote Evaluation Missions*, World Bank, Washington, D. C.

World Bank (2020b), *Summary of Lessons from the Knowledge Series on Using Technologies and Tools for Remote Data Collection: Experiences from Evaluation Offices of Multilateral Development Banks*, World Bank, Washington, D. C.

7. ANNEX

7.1 Classification of the findings for the official implementing organisations

This evaluation report presents the findings, conclusions and recommendations in relation to the involved organisations' requirements for applying the OECD-DAC and/or the DeGEval standards, as well as organisation-specific quality standards, plus the quality criteria of the sustainability meta-evaluation (chapters 4 and 5). Below, the findings for the official implementing organisations – the BGR, GIZ, KfW and PTB – are analysed and shown separately.

Unlike NGOs, the official implementing organisations are required to implement the BMZ Evaluation Policy. Based on the primacy of "different instruments with different roles" in bilateral official development cooperation, official implementing organisations must implement the BMZ Evaluation Policy indirectly. For NGOs, on the other hand, this policy provides guidance. Accordingly, the policy (BMZ, 2021: 21) states: "This policy is binding for the BMZ and the official implementing organisations (BGR, GIZ, KfW and PTB) […] It […] provides guidance for German civil society organisations – in each case in conjunction with contractual agreements, administrative regulations or funding guidelines of the BMZ with respect to these organisations." The findings, conclusions and recommendations are therefore of particular interest to the official implementing organisations and the BMZ. For this reason, the specific findings for the official implementing organisations are also shown separately here (Box 9).

Box 9 General conclusion on the understanding of quality and the application of the OECD-DAC and DeGEval quality standards by the implementing organisations

The focus on application of the OECD-DAC and DeGEval quality standards by the four official implementing organisations revealed a positive picture. The BGR, GIZ, KfW and PTB each applied the quality criteria between 61 and 87 per cent of the time. Application was rated as "largely achieved" for three official implementing organisations. The GIZ was the only organisation to obtain an average score of 87 per cent, thus fully achieving application.

Application of the quality criteria differed, sometimes markedly, both within the official implementing organisations and between them. One point of criticism is that not all quality criteria and their application were listed in organisational documents. Nor – in the case of intentional non-application – was this explained. GIZ is so far the only organisation that refers to the application of individual (though not all) quality standards in its internal organisational documents, explains this, and links it to the standards documents of the OECD-DAC and DeGEval. Furthermore, non-applications at the level of the evaluation were almost never listed by all organisations.

What understanding of evaluation quality do the involved German development cooperation organisations have? (Evaluation question 1)[86]

- The four official implementing organisations' understanding of quality was based on the OECD-DAC and DeGEval standards.
- The GIZ was the only official implementing organisation to also cite organisation-specific quality standards that went beyond the internationally recognised quality standards.

86 The findings, conclusions and recommendations for the GIZ and KfW in relation to evaluation question 2c (To what extent are strengths and weaknesses evident in the application of the quality criteria of the sustainability meta-evaluation in the evaluations of the GIZ and KfW?) are presented in section 4.2.4 and chapter 5, and are not discussed again here. Evaluation question 3 (To what extent are country-specific, evaluation-specific and organisation-specific factors linked to the application of quality standards?) was not examined separately for the official implementing organisations, as no divergent causal relationships between the examined explanatory factors and quality standards were assumed for these organisations. Accordingly, no further findings are included here.

To what extent are strengths and weaknesses evident in the application of the OECD-DAC and the DeGEval standards, and organisation-specific standards, in the evaluations of the involved German development cooperation organisations? (Evaluation question 2a and 2b)

a) Strengths were evident in the application of the quality criteria.

- The 37 quality criteria were applied on average 71 per cent of the time across all official implementing organisations (this is 3 per cent higher than the average for all organisations in group 1; section 4.2.1). The three most frequently applied quality criteria were "evaluation ethics (22)", the average "application of the OECD-DAC criteria (33-37)" and "description of the evaluation object (1)".
- On average, all official implementing organisations applied the quality criteria at least more than 60 per cent of the time, with GIZ being the only organisation to achieve them fully on average (87 per cent). For the other three organisations, on average this was largely the case (KfW and PTB 68 per cent each, BGR 61 per cent).
- All official implementing organisations had quality criteria whose application was rated as "not achieved", "barely achieved" and/or "partly achieved".
- The GIZ – which was the only official implementing organisation with organisation-specific quality standards – achieved these fully.

b) Weaknesses were found in the identification and systematic prescription of relevant quality criteria in the organisational documents, and the traceability of the application/non-application of some quality criteria at evaluation level.

- The three least applied quality criteria were "incorporation of evaluation capacity development (26)", "description of the methodological adequacy (7)" and "inclusion of partner-country evaluators (30)".
- The 15 quality criteria examined in the organisational documents were largely addressed at BGR (approx. 53 per cent) and GIZ (approx. 73 per cent), and partly addressed at KfW and PTB (approx. 33 per cent each). Accordingly, the remaining quality criteria were not identifiable in the organisational documents.
- In the organisational documents, no explicit reference was made to the standards documents when describing the quality criteria examined – with the exception of GIZ (for five out of 15 quality criteria). Evaluation processes for implementing the quality standards were also not always prescribed.
- A non-application of quality standards was only recorded in the organisational documents of the BGR and GIZ for the quality criterion "publication of the evaluation report (11)". An explained non-application at evaluation level was only documented in individual cases by one implementing organisation.

Understanding of quality among the official implementing organisations

The four official implementing organisations' understanding of quality was based on the OECD-DAC and/or the DeGEval standards. Only the GIZ had prescribed organisation-specific quality standards going beyond that. The BGR, GIZ, KfW and PTB required themselves to apply the DeGEval standards in their organisation-specific evaluation documents or through membership, and were required to apply the OECD-DAC standards through the BMZ Guidelines for bilateral Financial and Technical cooperation with cooperation partners of German development cooperation (BMZ, 2021b). For the BGR, KfW and PTB, the requirement to apply the OECD-DAC standards existed exclusively through the binding nature of these BMZ guidelines.

Rating of the application of the quality criteria

Application of the quality criteria averaged 71 per cent for all four official implementing organisations. For the GIZ it was 87 per cent, for KfW and PTB 68 per cent each and for BGR 61 per cent. A non-application of quality criteria, on the other hand, was not prescribed – beyond individual cases - at the evaluation level for any of the four official implementing organisations. This can be rated as "barely achieved". A detailed analysis of the application of the 37 quality criteria by organisation can be found in Table 7. Overall, approximately 8 per cent of the quality criteria were rated as "partly achieved" by GIZ (three out of 31 quality criteria), approximately 26 per cent each by KfW and PTB (ten out of 31) and 43 per cent partly achieved or worse by BGR (16 out of 31); Table 8).

Table 7 Overview of the findings on the official implementing organisations

QC	QS	SC	Level	Quality criterion	BGR	GIZ	KfW	PTB	Average
1	1	R & M	E	Evaluation object	93 %	100 %	98 %	96 %	97 %
/	2	R & M	E	Context	58 %	75 %	75 %	76 %	71 %
2	2a	R & M	E	Context of development intervention	76 %	72 %	97 %	85 %	82 %
3	2b	R & M	E	Context of findings	41 %	77 %	54 %	67 %	60 %
4	3	R & M	E	ToC	41 %	87 %	38 %	43 %	52 %
5	4	R & M	E	Evaluation interest	67 %	99 %	43 %	81 %	72 %
6	5	R & M	E	Information sources	83 %	98 %	77 %	79 %	84 %
7	6	R & M	E	Description of methodology	2 %	66 %	20 %	37 %	31 %
8	7	R & M	E	Existence of IR	0 %	100 %	87 %	100 %	72 %
9	8	R & M	E	Coherence	63 %	96 %	52 %	61 %	68 %
10	9	R & M	E	ToRs	63 %	65 %	4 %	56 %	47 %
/	10	U	Mi	Accessibility	50 %	100 %	60 %	75 %	71 %
12	10a	U	O	Publication report	0 %	100 %	100 %	0 %	50 %
11	10b	U	O	Publication summary	0 %	100 %	100 %	100 %	75 %
13	10c	U	E	Accessibility stakeholders	100 %	100 %	82 %	100 %	96 %
14	11	U	O	Competence	75 %	100 %	75 %	100 %	88 %
15	12	U	O	Timeliness	50 %	75 %	75 %	75 %	69 %
/	13	U	E	Recommendations	48 %	71 %	36 %	83 %	59 %
16	13a	U	E	Addressees recommendations	24 %	72 %	18 %	99 %	53 %

QC	QS	SC	Level	Quality criterion	BGR	GIZ	KfW	PTB	Average
17	13b	U	E	Actionable recommendations	72 %	70 %	54 %	67 %	66 %
18	14	U	O	Evaluation efficiency	50 %	100 %	100 %	75 %	81 %
19	15	P, I & F	O	Differences of opinion	100 %	75 %	/	0 %	58 %
/	16	P, I & F	Mi	Stakeholder involvement	87 %	98 %	12 %	69 %	67 %
20	16a	P, I & F	E	Involvement stakeholders (internal/external)	87 %	98 %	12 %	69 %	67 %
21	16b	P, I & F	O	Involvement partners*	100 %	100 %	100 %	50 %	88 %
22	17	P, I & F	O	Ethics	/	100 %	/	100 %	100 %
/	18	P, I & F	Mi	Independence	63 %	91 %	79 %	14 %	62 %
23	18a	P, I & F	E	Description organisational independence	44 %	87 %	69%	21 %	55 %
24	18b	P, I & F	O	Description impartiality	100 %	100 %	100 %	0 %	75 %
25	19	R & M	E	Executive summary	89 %	76 %	44 %	89 %	75 %
26	20	P, I & F	E	Evaluation capacity development	24 %	75 %	7 %	20 %	32 %
27	21	P, I & F	O	Joint evaluation	25 %	100 %	100 %	25 %	63 %
28	22	P, I & F	O	Partner orientation	25 %	50 %	50 %	75 %	50 %
/	23	P, I & F	Mi	Evaluation team	20 %	61 %	/	27 %	36 %
29	23a	P, I & F	O	Gender balance	50 %	50 %	/	75 %	58 %
30	23b	P, I & F	E	Partner-country evaluators	6 %	66 %	12 %	4 %	22 %
31	24	U	O	Resources	100 %	100 %	75 %	75 %	88 %
32	25	U	O	Management response	50 %	50 %	75 %	100 %	69 %
/	26	N	E	OECD-DAC criteria	98 %	100 %	100 %	100 %	99 %
33	26a	N	E	Relevance criterion	100 %	100 %	100 %	100 %	100 %
34	26b	N	E	Effectiveness criterion	100 %	100 %	100 %	100 %	100 %
35	26c	N	E	Efficiency criterion	89 %	100 %	100 %	100 %	97 %

QC	QS	SC	Level	Quality criterion	BGR	GIZ	KfW	PTB	Average
36	26d	N	E	Impact criterion	100 %	100 %	100 %	100 %	100 %
37	26e	N	E	Sustainability criterion	100 %	100 %	100 %	100 %	100 %
Average					**61 %**	**87 %**	**68 %**	**68 %**	**71 %**

Source: DEval, authors' own table

*Note: QC = quality criterion; QS = quality standard; SC = standard cluster; E = evaluation level; O = organisational level; Mi = mixed; R & M = reporting and methods; P, I & F = participation, independence and fairness; U = usefulness; N = no standard cluster. * This quality criterion was not included in the calculation for the quality standard "stakeholder involvement" (see section 4.2.1).*

Table 8 shows the findings as numbers and percentages for each quality criterion and rating category.

Table 8 Numbers and percentage values for quality criteria by category and organisation

	BGR		GIZ		KfW		PTB	
Category	# QC	% QC	# QC	% QC	# QC	% QC	# QC	% QC
Not achieved	3	8 %	0	0 %	0	0 %	3	8 %
Barely achieved (> 0 % ≤ 25 %)	6	16 %	0	0 %	6	16 %	4	11 %
Partly achieved (> 25 % ≤ 50 %)	7	19 %	3	8 %	4	11 %	3	8 %
Largely achieved (> 50 % ≤ 75 %)	5	14 %	9	24 %	8	22 %	10	27 %
Achieved (> 75 % ≤ 99 %)	6	16 %	8	22 %	5	14 %	6	16 %
Exceeded (100 %)	9	24 %	17	46 %	11	30 %	11	30 %
No information in online survey[a]	1	3 %	0	0 %	3	8 %	0	0 %
Total	**37**	**100 %**	**37**	**100 %**	**37**	**100 %**	**37**	**100 %**

Source: DEval, authors' own table

Note: QC = quality criterion. At the GIZ, all three organisation-specific quality criteria were achieved. [a] In contrast to the document analysis, the online survey gave the organisations the option of not providing any information on the application of specific quality criteria, without this being counted as "barely applied". This enabled the responsible officers of the evaluation units/desks to avoid having to make any inappropriate assessment of the application of a quality criterion across all evaluations.

The systematic inclusion of the application of the quality standards in the organisational documents was rated as "partly achieved" for KfW and PTB and as "largely achieved" for BGR and GIZ. Systematic inclusion meant that each quality standard was incorporated in organisational documents and the respective processes required for application were described, so that the application of all relevant quality standards can be ensured at evaluation level. This information can then be used by the various actors involved in designing, conducting and writing up the evaluation. Systematic institutionalisation was investigated for 15 of the quality criteria examined. This revealed that for up to ten quality criteria there was no institutionalisation in the organisational documents (BGR: N = 7; GIZ: N = 4; KfW: N = 10; PTB: N = 10). Furthermore, only two implementing organisations described a non-application of the quality criterion "publication of the evaluation report (11)", and only GIZ added a reference to the standards documents for five quality criteria (BGR: N = 0; GIZ: N = 5; KfW: N = 0; PTB: N = 0; Table 9). Overall, the GIZ showed the best institutionalisation of the quality standards in its organisational documents.

Table 9 Documentation of the application/non-application of selected quality criteria in the organisational documents of the official implementing organisations

	BGR	GIZ	KfW	PTB
Quality assurance with inception report (8)	A	A + E + R	A	A
Publication of the evaluation report (11)	NA	A + E	NA	none
Publication of the executive summary (12)	A	A + E	A	none
Competence of the evaluators (14)	A	A + E + R	A	A + E
Timeliness of the findings (15)	none	A + E + R	none	none
Evaluation efficiency (18)	none	A + E + R	none	none
Transparency of differences of opinion (19)	none	none	none	none
Involvement of partners (21)	A	A	A	none
Evaluation ethics (22)	A	A + E + R	none	A + E
Description of the impartiality of the evaluators (24)	A	none	none	A
Consideration of joint evaluations (27)	none	A	none	none
Partner-country orientation (28)	none	none	none	none
Gender balance in the evaluation team (29)	none	none	none	none
Sufficient resources available (31)	none	A	none	none
Existence of a management response (32)	A	A	none	A

Source: DEval, authors' own table

Note: none = no information supplied on application of the quality criterion; A = information supplied on application of the quality criterion; NA = information supplied on non-application of the quality criterion; A + E = information supplied on application of the quality criterion, with explanation; A + E + R = information supplied on application of the quality criterion, with explanation and with reference to the standards document

The recommendations made for all organisations (chapter 5) also apply to the four official implementing organisations; all that differs between BGR, GIZ, KfW and PTB is the extent of revision of their own evaluation practice. Since the application of quality standards is a basic element of evaluation practice and no quality standard has been given priority in the standards documents so far, the urgency of implementing recommendations 1 to 3[87] is high for all four implementing organisations. Recommendation 4 can be dealt with as a lower priority, as it represents an option – albeit an important

87 Recommendations 1 (identification and systematic prescription of application at organisational level), 2 (identification and systematic prescription of non-application at organisational level) and 3 (application of quality standards at evaluation level).

one – to improve the application of quality standards. As recommendation 5 complements recommendations 1 and 3, there is no need for GIZ and KfW to act until corresponding quality criteria of the sustainability meta-evaluation and further quality criteria of the BMZ Evaluation Policy have been included in a BMZ analysis grid.

7.2 List of quality criteria

Table 10 presents the quality criteria examined. It shows both the numbering assigned to them, and the long version of their names. Each quality criterion also has a short name, which is used for example in the tables and graphs etc. A list of the short names can also be found in section 4.1.1 of the online appendix.

Table 10 Overview of the numbering and names of the quality criteria examined

No. of QC	Name of QC (name of QS)	No. of QC	Name of QC (name of QS)
1	Description of the evaluation object	20	Involvement of internal and external stakeholders (stakeholder involvement)
2	Description of the context of the development intervention (consideration of the context)	21	Involvement of partners (stakeholder involvement)
3	Incorporation of the context of the findings (consideration of the context)	22	Evaluation ethics
4	Description of the Theory of Change	23	Description of the organisational independence of the evaluators (independence of the evaluators)
5	Description of the evaluation interest	24	Description of the impartiality of the evaluators (independence of the evaluators)
6	Clarity of the information sources	25	Information content of the executive summary
7	Description of the methodological adequacy	26	Incorporation of evaluation capacity development
8	Quality assurance with inception report	27	Consideration of joint evaluations
9	Coherence of data-findings-conclusions	28	Partner-country orientation
10	Information content of the terms of reference	29	Gender balance in the evaluation team (composition of the evaluation team)
11	Publication of the evaluation report (accessibility)	30	Inclusion of partner-country evaluators (composition of the evaluation team)
12	Publication of the executive summary (accessibility)	31	Sufficient resources available

No. of QC	Name of QC (name of QS)	No. of QC	Name of QC (name of QS)
13	Accessibility for stakeholders (accessibility)	32	Existence of a management response
14	Competence of the evaluators	33	Application of the OECD-DAC criterion – relevance (application of the OECD-DAC criteria)
15	Timeliness of the findings	34	Application of the OECD-DAC criterion – effectiveness (application of the OECD-DAC criteria)
16	Addressees of the recommendations (usefulness of the recommendations)	35	Application of the OECD-DAC criterion – efficiency (application of the OECD-DAC criteria)
17	Actionable recommendations (usefulness of the recommendations	36	Application of the OECD-DAC criterion – impact (application of the OECD-DAC criteria)
18	Evaluation efficiency	37	Application of OECD-DAC criterion – sustainability (application of OECD-DAC criteria)
19	Transparency of differences of opinion		

Source: DEval, authors' own table
Note: QC = quality criterion; QS = quality standard.

7.3 Rating scale for DEval evaluations

Category	Meaning
Exceeded	Findings demonstrate an application of the quality criterion that exceeds the benchmark.
Achieved	Findings demonstrate an application of the quality criterion that meets the benchmark.
Largely achieved	Findings predominate which demonstrate an application of the quality criterion that meets the benchmark.
Partly achieved	There are some findings which demonstrate an application of the quality criterion that meets the benchmark.
Barely achieved	Barely any findings demonstrate an application of the quality criterion that meets the benchmark.
Not achieved	There are no findings which demonstrate an application of the quality criterion that meets the benchmark.

7.4 Evaluation matrix

	Organisation-specific data and documents	Scientific and empirical literature	Interviews	Focus group discussions
Evaluation question 1	X		X	
Evaluation question 2a	X		X	
Evaluation question 2b	X			
Evaluation question 2c	X			
Evaluation question 3	X	X		X

Source: DEval, authors' own table

7.5 Timeline of the evaluation

Time frame	Tasks
08/2020	Memo sent
12/2020	Reference group meeting to discuss concept note
12/2020 – 04/2021	Inception report written
04/2021	Reference group meeting to discuss inception report
04/2021 – 04/2022	Data collection and synthesis of the findings
04/2022	Reference group meeting to discuss the findings
04/2022 – 08/2022	Evaluation report written
08/2022	Reference group meeting to discuss the conclusions and recommendations
12/2022	Completion of evaluation after layout

7.6 Evaluation team and contributors

Core Team	
Dr Kerstin Guffler	Team leader
Dr Nico Herforth	Team leader (interim: 02/2021 – 02/2022)
Laura Kunert	Evaluator
Marian Wittenberg	Evaluator
Rebecca Maicher	Project administrator
Contributors	
Dr Martin Noltze	Internal peer reviewer
Prof. Dr Wolfgang Beywl	External peer reviewer
Prof. Dr Thomas Widmer	External evaluation quality expert
Christian Süper	Student employee
Rayan Doukali	Student employee
Maria Villa-Guillen	Student employee
Lucia Citoler	Student employee
Annika Grotrian	Student employee
Responsible	
Amélie Gräfin zu Eulenburg	Head of Department