SUSTAINABILITY IN GERMAN DEVELOPMENT COOPERATION

Meta-evaluation
2018

DEval
GERMAN INSTITUTE FOR DEVELOPMENT EVALUATION

This meta-evaluation ‘Sustainability in German development cooperation’ is part of DEval’s thematic focus on sustainability. The meta-evaluationis complemented by an accompanying evaluation synthesis. Linked by an integrated evaluation design, the two reports share a common database and pursue complementary objectives.

	Meta-evaluation	Evaluation synthesis
Aims	Analyse the practice of evaluating the sustainability of German development cooperation projects to date Reconstruct the understanding of sustainability in German development cooperation to date, and compare this with the modern understanding inherent in the 2030 Agenda for sustainable development Support the design of evaluation practices that are in conformity with the 2030 Agenda	Analyse the factors affecting the rating of project sustainability Study the sustainability rating of German development cooperation projects Highlight ways of increasing the sustainability of German development cooperation projects Support the strategic and operational alignment of German development cooperation with the requirements of the 2030 Agenda for Sustainable Development
Methods	Systematic quality analysis and quantitative content analysis	Multivariate regression analysis
Database	Evaluation reports on German development cooperation projects plus secondary data	
Integrated design	The findings of the quantitative content analysis performed in the meta-evaluation were integrated into the regression analyses of the evaluation synthesis as explanatory variables. The findings of the qualitative analysis performed by the meta-evaluation were integrated into the regression analyses of the evaluation synthesis as a weighting factor for the explanatory value of the observations.	

SUSTAINABILITY IN GERMAN DEVELOPMENT COOPERATION

Meta-evaluation
2018

Imprint

Published by
German Institute for Development Evaluation (DEval)
Fritz-Schäfer-Straße 26
53113 Bonn, Germany

Tel: +49 (0)228 33 69 07-0
Email: *info@DEval.org*
www.DEval.org

Authors
Dr. Martin Noltze
Dr. Michael Euler
Ida Verspohl

Responsible
Prof. Dr. Jörg Faust (until June 2016)
Dr. Sven Harten (since June 2016)

Design
MedienMélange:Kommunikation!, Hamburg
www.medienmelange.de

Translation
Dr. John Cochrane

Photo credits
Gui Yongnian/123rf.com (Cover), Olaf Speier/Alamy Stock Foto (Chap. 1), dbimages/Alamy Stock Foto (Chap. 2 + 3), Dzianis Apolka/Alamy Stock Foto (Chap. 4), imageBROKER/Alamy Stock Foto (Chap. 5), Riccardo Lennart Niels Mayer/ 123rf.com (Chap. 6), Oleksandr Roslyak/123rf.com (Chap. 7)

Bibliographical reference
Noltze, M., M. Euler and I. Verspohl (2018), Meta-evaluation of sustainability in German development cooperation, German Institute for Development Evaluation (DEval), Bonn

Printing
DCM Druck Center
Meckenheim

MIX
Paper from responsible sources
FSC® C081774

ISBN 978-3-96126-069-0 (print)
ISBN 978-3-96126-070-6 (PDF)

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 cooperation.

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:
www.deval.org/en/evaluation-reports.html

Requests for print copies of this report should be sent to:
info@DEval.org

Acknowledgements

In its work on this report, the evaluation team was supported by a large number of individuals and organisations. We would like to express our cordial thanks to all of them.

First of all, the support provided by the reference group was key to the success of this meta-evaluation and the accompanying evaluation synthesis. In this connection we would also like to say a special word of thanks to the divisions of the German Federal Ministry for Economic Cooperation and Development (BMZ) involved, especially Division 105 (Michaela Zintl, Katrin von der Mosel and Berthold Hoffman) and Division 300 (Gottfried von Gemmingen-Guttenberg, Dr. Ingolf Dietrich, Dr. Maya Schmaljohann, Cormac Ebken and Ruben Werchan), as well as the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ; Dr. Ricardo Gómez, Dorothea Giesen-Thole, Valentin Dyckerhoff, Katrin Ladwig and Cornelia Skokov), and KfW Development Bank (KfW; Prof. Dr. Eva Terberger, Martin Dorschel, Thomas Gietzen and Christian Schönhofen). In particular we would like to thank them for their many suggestions and comments in what was an open and discerning discussion process. The GIZ and KfW deserve our special thanks for their strong support when collecting the data – without the extensive data and documents which they provided, we would not have been able to perform our evaluation work.

We would also like to thank our colleagues at DEval, who kept us in good spirits and provided critical support to the evaluation process. Here we owe a particular debt of thanks to our in-house DEval peer reviewers Dr. Kerstin Guffler and Solveig Gleser, and our Director Prof. Dr. Jörg Faust, for their many suggestions and comments. We are also grateful to Thomas Wencker for his critical perspective and constructive proposals. Our thanks are also due to Cornelia Michaels-Lampo and our other administrative staff for the support they provided during the evaluation work. Our in-house Media and Public Relations Unit and the report's translator also deserve a special thank you.

We would also like to express our gratitude to Jana Preiß, who helped us carry out the contextual study for the meta-evaluation as part of her associate master's thesis.

Our interns and student assistants Helena Heberer, Niklas Witzig, Grisel Orozco, Sarah Stahlmann and Lea Smidt, whose support made a valuable contribution to the success of the evaluation, also deserve our gratitude. We would like to sincerely thank them for their huge commitment and personal dedication.

A special word of thanks is also owed to our external peer reviewer Prof. Dr. Sebastian Vollmer. His numerous inputs on content and methodology made a crucial contribution to the quality of these evaluation reports.

Finally, we would like to thank our colleagues at the Competence Centre for Evaluation Methodology, who were there to support us throughout the evaluation process with searching questions and suggestions for our methodology.

EXECUTIVE SUMMARY

Background, purpose and object of the evaluation

The 2030 Agenda for Sustainable Development emphasises the global significance of the sustainability principle. Sustainability is thus now defined in relation to key principles of sustainable development. Universality, shared responsibility and accountability, synergy between social, economic and environmental development, and inclusiveness, form the principles of the modern understanding of sustainable development.

Germany has committed to the principles of the 2030 Agenda and pledged to implement them in its development cooperation. Within the German development cooperation system, the notion of sustainability has for some time been an integral part of the development debate. A basic distinction is drawn here between 'sustainable development' and 'the continuation of development results over time'. To what extent these two aspects are reflected in or correspond to the modern understanding of sustainability after the 2030 Agenda still remains an open question. So far, neither the conceptual understanding of sustainability nor the way it is dealt with in practice in German development cooperation has been subjected to systematic analysis. The current development agenda now provides the occasion for a comprehensive study of sustainability, which has been the guiding principle of German development cooperation for many years.

The purpose of the present meta-evaluation is to undertake a first comprehensive and systematic survey of the practice of evaluating sustainability in German development cooperation. This empirical study of existing practice is designed to reconstruct the understanding of sustainability in German development cooperation, which has to date been somewhat difficult to pin down, and then compare this with the modern understanding of sustainability based on the principles of the 2030 Agenda. In other words, the purpose of the meta-evaluation is to support the design of evaluation practices that conform to the 2030 Agenda.

The object of the meta-evaluation is how practitioners actually assess sustainability in German development cooperation projects, as reflected in the evaluation reports of Germany's two major official implementing organisations – the KfW Development Bank (KfW), and the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. Both implementing organisations assess the sustainability of projects using the international evaluation criteria of the Development Assistance Committee of the Organisation for Economic Co-operation and Development (OECD). Based on a guideline published by the German Federal Ministry for Economic Cooperation and Development (BMZ) in 2006, the continuation of development results over time forms the core of the evaluation criterion 'sustainability'. Furthermore, when the meta-evaluation began the team proceeded on the assumption that the notion of results – in conjunction with the other evaluation criteria (relevance, effectiveness, efficiency and impact) – also implies sustainable development.

Methodology

This study is a thematic meta-evaluation. In this case the traditional meta-evaluation design involving a purely qualitative assessment was extended to include a systematic examination of 'sustainability' when used as a criterion to assess development cooperation. The database for the meta-evaluation comprised a representative random sample of 513 evaluation reports on German Technical and Financial Cooperation projects. As part of an integrated research design, the findings of the meta-evaluation were also fed into the accompanying evaluation synthesis, which examines the factors affecting sustainability.

Key findings, conclusions and recommendations concerning the assessment of sustainability in German development cooperation

The findings of the present meta-evaluation confirm the prior assumption that the evaluation criteria imply not only the continuation of development results over time, but also sustainable development. Hence these findings demonstrate empirically for the first time that in the evaluation of German development cooperation, sustainability is already being understood in a comprehensive sense, and evaluated and assessed accordingly. At the same time a significant discrepancy exists in relation to the aspirations of the 2030 Agenda. Key principles of the 2030 Agenda, such as synergy between the dimensions of sustainability, are not yet a systematic element of assessment practice. The findings thus

refute the possible assumption that the DAC evaluation criteria are based exclusively on a narrow understanding of sustainability that would be confined to the continuation of results. Nevertheless, they do point to significant discrepancies in relation to the modern understanding of sustainability inherent in the 2030 Agenda.

The findings also demonstrate that in practice, sustainability is currently being assessed unsystematically and inconsistently due to the absence of a conceptual framework for a comprehensive understanding of sustainability. The key questions proposed in the BMZ guideline in 2006 are also not being applied systematically. Overall, it is evident that the DAC criteria as they stand do permit the evaluation of sustainability understood in a comprehensive sense, but by no means prescribe this on a systematic and binding basis. This lack of a systematic approach means that the value of aggregating the sustainability score across different projects is limited by the inherent lack of comparability between the scores for the individual projects, which is not conducive to learning from evaluations. At present, a rigorous comparison of the sustainability of projects is only possible at considerable expense and with considerable effort – such as the effort made in preparing the present expanded meta-evaluation and the accompanying evaluation synthesis.

In the future, working with the 2030 Agenda and the sustainability of development cooperation projects in evaluations will be a global task. With respect to German development cooperation, this meta-evaluation has identified a specific need for action. The conclusions call for a reform of existing evaluation practices. Alongside the idea of harmonisation and coordination contained in the Paris Declaration and the Accra Agenda for Action, the universal nature of the 2030 Agenda also calls for sharing and coordination at the international level. The recommendations below are designed to support the ongoing reform process at the level of German development cooperation, and enrich the debates at the international level. First of all the authors present their key recommendations for further developing the practice of evaluation. These are then followed by basic recommendations for further developing the evaluation system.

Recommendations on further developing evaluation practice

The evaluation team recommends that in the future the BMZ and the implementing organisations should evaluate the sustainability of projects based on the principles of the 2030 Agenda for Sustainable Development, within the framework of an additional assessment criterion.

As well as including sustainability as conceptualised in the 2030 Agenda as an additional criterion, the BMZ should sharpen the conceptual focus of the DAC criteria and make the BMZ guidelines for applying the DAC criteria more binding.

As part of the reform of evaluation criteria for assessing the performance of development cooperation projects, the evaluation team recommends that the BMZ retain the existing OECD-DAC criterion of sustainability – understood as implying the continuation of results – and align its key questions with this element.

With respect to the principles of the 2030 Agenda, the GIZ and KfW should investigate how in future evaluations they can identify and assess the unintended effects of a project and the interactions between the dimensions of sustainability.

The implementation and conceptual elaboration of the recommendations on evaluation practice should take place in Germany on the basis of a joint process led by the BMZ and involving the implementing organisations and DEval. The team recommends that this process, including a pilot phase, should be completed by the end of 2018, in order to guarantee from 2019 onwards that evaluation in German development cooperation is in conformity with the 2030 Agenda. At the same time the ongoing reform process within the German development cooperation system should be reviewed with regard to its international connectivity, and discussed in the appropriate forums.

Recommendations on further developing the evaluation system

The evaluation team recommends that the BMZ develop an overarching evaluation strategy that in the course of time sets thematic priorities.

In the evaluation strategy the BMZ should define what requirements arise from the questions raised by the 2030 Agenda for the various evaluations – i.e. at the level of modules, programmes and country strategies.

Key findings, conclusions and recommendations concerning the quality of evaluations in practice

The meta-evaluation analysed not only the assessment of sustainability in German development cooperation, but also evaluation quality. The findings of the quality analysis provide an indication of the robustness of the findings and conclusions of the evaluations concerning the sustainability of German development cooperation.

They demonstrate that the excellent quality of the findings and conclusions obtained by the GIZ and KfW from their module evaluations is appropriate for evaluations of that size. As well as describing the object of the evaluation, most of the reports include a logical description of the causal links to be analysed and the methodological approach. German development cooperation is characterised by a high degree of coverage by evaluations. The GIZ submits almost all modules to a systematic evaluation of results, while the KfW operates with a representative random sample.

However, it also emerged that the quality of evaluations at module level can be improved. Systematic methods of analysis and triangulation should be used to increase efforts to detect causal relationships. The same thing applies to the plausibility of findings and conclusions in the evaluation reports. It is also important to focus the available resources on the purpose of the evaluation. In decentralised evaluations, evaluators have so far set out not only to evaluate as such, but also to appraise. Furthermore, results and sustainability can be substantiated by selecting an appropriate point in time at which to conduct the evaluation. Ex-post evaluations offer an opportunity to actually observe results and their sustainability after a certain interval following completion of the project. The decentralised evaluations conducted during the course of a project, on the other hand, substantiate sustainability purely on the basis of an assessment of future likelihood. Given the limited availability of data in the context of development cooperation, monitoring data are an important source. However, their potential for reliably substantiating results and sustainability is not yet being utilised to the full.

The findings of the meta-evaluation also revealed an interesting link between the quality of evaluations and the quantity of information produced. As the quality of evaluations rises, so too does the number of criteria applied to assess sustainability. More sophisticated evaluations place the assessment of sustainability on a broader footing, and are conducive to the generation of reliable findings. There is no direct link between the quality of evaluation and the assessment of an individual criterion or the overall assessment of the sustainability of a project.

Given the link between quality and the detail in which sustainability is dealt with in evaluations, plus the close link between substantiating results and substantiating sustainability, a number of recommendations arise in relation to the quality of evaluations and the underlying evaluation system. Here too the authors will first of all present recommendations for further developing evaluation practice. These are then followed by recommendations on further developing the evaluation system.

Recommendations on further developing 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.

Bearing in mind the low importance persistently ascribed to monitoring data in module evaluations, the implementing organisations should systematically examine what obstacles exist here and how these can be overcome. In this context they should examine whether project monitoring systems can be linked through their objectives systems to the system of goals and targets that make up the Sustainable Development Goals (SDGs).

To ensure transparency and incentivise clear reporting the GIZ and KfW should, while remaining mindful of the opportunities and risks, explore the possibility of publishing their evaluation reports in full – perhaps initially in a pilot phase – and informing the BMZ of the lessons they learn in the process.

To raise the quality of evaluation, the team recommends that GIZ institutionalise the role of quality assurance in the Evaluation Unit on a long-term basis. In the future, all module evaluations should be managed by the Unit.

To help raise evaluation quality, appraisal and evaluation should be separated at the GIZ.

Regarding the appropriate point in time at which to reliably substantiate results and sustainability, greater importance should once again be attached to ex post evaluations. When ex post evaluations are being conducted, both the GIZ and KfW should ensure that the importance of management is understood. This can involve for instance defining key focuses, or selecting an appropriate point in time for the evaluation.

Recommendations on further developing the evaluation system

To promote joint learning and accountability, the team recommends that the BMZ harmonise the practice of evaluation by the GIZ and KfW on the basis of the joint procedural reform (*Gemeinsame Verfahrensreform*, GVR) and the Guidelines for bilateral Financial and Technical Cooperation. In this context the BMZ should issue firm instructions concerning the timing, scope and rating system in order to standardise the types of evaluation for module evaluations.

By defining uniform minimum standards the BMZ should support the exhaustive use of potential to raise evaluation quality in module evaluations.

The BMZ should require the implementing organisations to make their evaluation reports clear and easy to understand, so that they can be read on a stand-alone basis. Depending on the outcome of a corresponding review, the BMZ should require the implementing organisations to publish their evaluation reports in full.

The BMZ should ensure that, in addition to the quality assurance of the module evaluations performed by the evaluation units of the GIZ and KfW, an external, cross-organisational meta-evaluation of a random sample of evaluations should be performed on a regular basis.

CONTENTS

Figures

Tables

ABBREVIATIONS AND ACRONYMS

BMZ
German Federal Ministry for Economic Cooperation and Development

DAC
Development Assistance Committee of the Organisation for Economic Co-operation and Development

FC
Financial Cooperation

GIZ
Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

GVR
Joint procedural reform (Gemeinsame Verfahrensreform)

KfW
KfW Development Bank

OECD
Organisation for Economic Co-operation and Development

OO
Overarching Objective

PE
Project Evaluation

PPR
Project Progress Review

SDGs
Sustainable Development Goals

TC
Technical Cooperation

1.

INTRODUCTION

This rigorous meta-evaluation represents a first comprehensive and systematic empirical analysis of the practice of evaluating and assessing the sustainability of German bilateral development cooperation projects. It is based on evaluations performed by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH and the KfW Development Bank (KfW) on projects financed through public funds of the German Federal Ministry for Economic Cooperation and Development (BMZ).

1.1 Background

The 2030 Agenda for Sustainable Development has given the principle of sustainability global significance. This strong emphasis on the notion of 'sustainability' is the consequence of a long-standing discussion in the international development debate, which was initiated by the United Nations in the 1980s and subsequently continued through various global development conferences. More recently, this debate culminated in the introduction of the 2030 Agenda for Sustainable Development. The continuing debate on sustainability represents nothing less than an engagement with the vital issue of the future viability of human and environmental development. The principle of sustainability is being emphasised in all quarters as pivotal to development. At the same time, the conceptual understandings underlying the term are comprehensive and complex.

The multidimensionality of the concept of sustainability is also reflected in development cooperation. Here a distinction is commonly drawn between 'sustainable development' and 'the continuation of development results over time'. This distinction does not, however, provide a conceptual clarification of the term **sustainability**. Ultimately it remains unclear how the term is actually being understood in practice in the policy field of development cooperation. However, the increased importance of the principle of sustainability resulting from the 2030 Agenda means that such imprecision can no longer be accepted. A comprehensive analysis of the understanding of sustainability is absolutely imperative. What is 'sustainability' understood to mean? How can sustainability be measured and assessed? How reliable is existing knowledge? These questions cannot be answered through theory alone. They also require a sound empirical analysis of this long-standing guiding principle of development cooperation. Taking an approach that is an open as possible, this meta-evaluation takes a comprehensive look at sustainability that is free from preconceptions. Where necessary, the approach allows scope for distinguishing between sustainable development and the continuation of development results. The background to these two aspects of sustainability is outlined briefly below and subsequently discussed at various points in the report.

In the international debate **sustainable development**, understood as part of the principle of sustainability, has a long history. As early as the 17th century, sustainability was emphasised in forest management as a guiding principle for the sound use of natural resources. According to this principle, foresters should only ever cut down as many trees as could grow back again using the available resources. More recently (in the 1970s), this basic principle was picked up in the debate on the 'Limits to [economic] growth' (Meadows et al., 1972). In the 1980s a multidimensional concept of social, economic and environmental sustainability then arose (Grunwald and Kopfmüller, 2006). Since the Brundtland Report was published in 1987, safeguarding the needs of future generations has also been at the heart of the idea of sustainability (World Commission on Environment and Development, 1987). Since the UN Conference on Environment and Development in Rio de Janeiro in 1992 this has been accepted internationally. Today, the 2030 Agenda for Sustainable Development is the logical consequence of an understanding of sustainability that is becoming increasingly integrated and complex. Universality, shared responsibility and accountability, inclusiveness, and synergy between social, economic and environmental development, are among the basic principles of the 2030 Agenda (UN, 2015). Furthermore, these principles are supported by 17 Sustainable Development Goals (SDGs) with 169 targets. In Germany, the relevance and the influence of the international debate on the conceptual understanding of sustainability as the guiding principle of development cooperation is undisputed (König and Thema, 2011). It remains unclear, however, to what extent development cooperation has in practice succeeded in integrating this increasingly complex

understanding of sustainability, or whether this is even possible. Sceptics assume that the degree of complexity goes beyond the capacities of development cooperation, and that the likelihood of achieving the goals associated with it is therefore diminishing continuously (Klasen, 2015; Nuscheler, 2007). This risk appears more relevant than ever, given the complexity of the 2030 Agenda. The way sustainability is dealt with in projects thus also typifies the tendency that development cooperation has to readily respond to complex challenges with complex solutions, which are then difficult to implement on the ground.

The second understanding of the concept of sustainability – based on the **continuation of results** – has also long been associated with development cooperation. This was emphasised in 1991 by the Development Assistance Committee (DAC) of the Organisation for Economic Co-operation and Development (OECD) as a criterion for measuring the performance of development cooperation projects (OECD, 1991). In 2006 the German Federal Ministry for Economic Cooperation and Development (BMZ) incorporated the OECD-DAC's understanding into its 'Evaluation criteria for German bilateral development cooperation. A guideline for evaluations performed by the BMZ and the implementing organisations' (BMZ, 2006). Since then, evaluations and appraisals have looked not only at relevance, effectiveness, efficiency and impact, but also sustainability. Sustainability is assessed with respect to three key aspects. First of all the durability of development results is assessed. The second aspect is the stability of the context in terms of social justice, economic performance, political stability and environmental balance. Thirdly, the risks and potential for the (continued) effectiveness of the project are assessed. On the basis of these three aspects, however, it becomes clear that the OECD-DAC understanding of sustainability is by no means confined purely to the continuation of development results, but is also closely linked to the concept of sustainable development.

Given the conceptual link between the aspects of sustainable development and the durability or continuation of results, this meta-evaluation proceeds on the assumption that in practice, sustainability is already being understood in a more comprehensive sense than the existing instructions and guidelines of the BMZ, GIZ and KfW would first lead the user to assume. In other words, the report assumes that a comprehensive understanding of sustainability has already become part of existing evaluation practice, and is reflected in it. However, the report also anticipates that the complexity of the understanding of sustainability and the lack of instructions have in the past led to sustainability being understood and assessed very inconsistently in evaluations.

1.2 Purpose of the meta-evaluation

This rigorous meta-evaluation is the first comprehensive and systematic empirical survey of the practice of evaluating sustainability in German development cooperation. It was prompted by the 2030 Agenda, through which the principle of sustainability has gained greater importance for development. The declared objective of the meta-evaluation is to survey the evaluation of sustainability in development cooperation. This empirical study of existing practice will thus facilitate a more concrete and detailed understanding of sustainability in German development cooperation, which has to date been somewhat difficult to pin down. Finally it will then also be possible to compare that understanding with the modern understanding of sustainability based on the principles of the 2030 Agenda. Accordingly, the key contribution of this meta-evaluation is twofold. First of all it will place the sustainability debate, which is often conducted on a purely theoretical level, on a broad empirical footing. Secondly, based on the findings it will develop a proposal on how sustainability should be evaluated in the future. Ultimately, the purpose of the meta-evaluation is to support the design of evaluation practices in conformity with the 2030 Agenda.

1.3 Object

The first, direct focus of the meta-evaluation is the practice of evaluating the sustainability of German development cooperation projects to date, as described in the evaluation reports of the implementing organisations. The second focus is the sustainability of German Financial and Technical Cooperation projects for development. Addressing the object

of the evaluation will allow a sound analysis of the understanding of sustainability in German development cooperation.

As the object of the evaluation is to be addressed as comprehensively as possible, the study is not restricted either to particular sectors, or to particular regions or types of project. As well as purely bilateral projects in specific countries, the study also covers regional, sectoral and global projects. To nevertheless guarantee the feasibility of this first rigorous thematic meta-evaluation, the object of the evaluation was narrowed down as follows.

First of all the analysis is confined to the practice of evaluation by the two major official implementing organisations – the KfW and GIZ.[1] Every year these two implementing organisations deliver a significant portion of public development finance, and each has a highly diversified portfolio of projects across all sectors and regions of German development cooperation. At the same time both implementing organisations have a high degree of evaluation coverage of individual projects (today referred to as modules). All evaluations assessed sustainability throughout.

The analysis was also narrowed down in terms of the period covered. The systematic and largely standardised assessment of sustainability as one of the criteria for the success of German development cooperation began in 2006 with the approval of the BMZ guideline on applying the DAC criteria. The analysis therefore includes only evaluations that were conducted and completed between July 2006 and the point at which the data were collected in October 2017.

1.4 Evaluation questions

The objectives of the evaluation were operationalised through five evaluation questions.

Evaluation question 1 – What criteria are used to assess sustainability in evaluations?

Evaluation question 2 – How appropriate is the practice of evaluation in German development cooperation as a means of assessing sustainability?

Evaluation question 3 – To what extent does the practice of evaluating sustainability in German development cooperation meet international standards and present-day demands?

Evaluation question 4 – What is the quality status of evaluation methods?

Evaluation question 5 – To what extent does the quality of evaluation methods affect the assessment of sustainability?

1.5 Structure of the evaluation report

The meta-evaluation report is structured as follows.

Chapter 2 begins by describing sustainability as a performance criterion in the aid effectiveness debate within German development cooperation (Section 2.1). Building on that, the conceptual framework of the meta-evaluation is then described (Section 2.2). The section concludes with a look at evaluation practices in German Technical and Financial Cooperation (Section 2.3).

The methodology of the meta-evaluation is described in Chapter 3. The chapter begins by describing the database (Section 3.1). It then details the methodology of the meta-evaluation with respect to the analysis of evaluation quality (Section 3.2) and assessment practice (Section 3.3). The methodology of the contextual study is contained in Section 3.4. The chapter is rounded off with a discussion of the limitations of the meta-evaluation (Section 3.5).

The findings of the analysis are presented in Chapter 4. The chapter begins with the findings on the quality of evaluations (Section 4.1), before moving on to the findings on the assessment of sustainability (Section 4.2), which are discussed in relation to the conceptual framework of the meta-evaluation.

[1] *Other official implementing organisations such as the Federal Institute for Geosciences and Natural Resources (BGR) and the Physikalisch-Technische Bundesanstalt (PTB) (Germany's national metrology institute) are not part of the analysis.*

Finally, Section 4.3 presents the findings on possible links between the quality of evaluations and assessment practice, and Section 4.4 presents the findings on the contextual study.

The conclusions and recommendations are contained in Chapter 5.

2.

THE EVALUATION OF SUSTAINABILITY IN GERMAN DEVELOPMENT COOPERATION

2.1
Sustainability in the aid effectiveness debate within German development cooperation

The international discourse on the sustainability principle that unfolded from the 1970s onwards pointed the way forward for the development of the understanding of sustainability in German development cooperation (see Section 1.1). However, there was a significant lag before practitioners actually translated that debate into an engagement with the sustainability of German development projects. Sustainability as an evaluation criterion did not become a focus of the German aid effectiveness debate until the end of the 1980s, for instance (Stockmann and Gaebe, 1993). At that time the understanding of sustainability embraced two aspects – sustainable development on the one hand, and the continuation of development results over time on the other.

The inclusion of sustainability in project evaluations as a criterion of performance was prompted in 1986 by a recommendation of the OECD-DAC, on the basis of which the BMZ later declared sustainability to be an important measure of the performance of German development cooperation. At the end of the 1980s sustainability was then included as a criterion in evaluations of official development cooperation, initially through the ex-post evaluations of the KfW (Stockmann and Gaebe, 1993). Later on, sustainability was also gradually incorporated into GIZ evaluations.

Finally, an aid effectiveness study commissioned by the BMZ in 1998/99 triggered a systematic engagement with sustainability as a performance criterion for development cooperation. The study examined long-term effectiveness in 32 selected ex-post evaluations of official German Technical and Financial Cooperation (TC and FC). In addition to these aggregate findings, an accompanying cross-section evaluation by Caspari subsequently focused on evaluation and assessment practices (2004). The debate which ensued made clear that at the time, sustainability was being understood and assessed in German development cooperation on a very heterogeneous basis, which placed considerable limitations on the scope for cross-section evaluation.

On the basis of these findings, and in the context of the OECD-DAC recommendations for harmonising the member states' evaluation systems, a working group led by the BMZ and involving the implementing organisations addressed the topic of 'joined-up evaluation'. The aim of this undertaking was to further standardise evaluations in bilateral German development cooperation and align them with international standards. Ultimately the work of this group led to the OECD-DAC evaluation criteria being made mandatory as a guiding framework for assessing the performance of German development cooperation (OECD, 1991). Since then, sustainability has been one of the binding evaluation criteria alongside relevance, effectiveness, efficiency and impact. The guideline operationalised sustainability as an evaluation criterion through three key questions (BMZ, 2006):

- To what extent are the positive changes generated by the development intervention and its results to be rated (summarily) as durable in relation to the development objectives?
- How stable is the context of the development intervention with respect to the factors 'social justice', 'economic performance', 'political stability' and 'ecological balance'?
- What risks and potentials are evident for the continued effectiveness of the development intervention, and how likely is it that these factors will materialise?

As already highlighted in Section 1.1, the underlying understanding of stability embraces both the aspect of durability and – via the notion of effectiveness/impact – the aspect of sustainable development. In other words, since 2006 the conceptual understanding of sustainability in German development cooperation has been comprehensive and complex (see Section 2.2). Consequently, only a synoptic look at the findings for the five evaluation criteria relevance, effectiveness, efficiency, impact and sustainability will then allow us to discuss and assess sustainability, understood conceptually as embracing both sustainable development and the continuation of results.

Once the BMZ had defined the key questions in its guideline of 2006, the GIZ and KfW then also went on to agree a binding rating scale. Since then, sustainability has been rated by

awarding one of four possible sustainability scores[2] that are included in each evaluation report: A score of 1 is awarded when the project's impact (which has so far been positive) is highly likely to continue unchanged or increase. A score of 2 is awarded when the project's impact (which has so far been positive) is highly likely to diminish only slightly. The score 3 means either that the impact (which has so far been positive) is highly likely to diminish significantly, but will remain positive, or that it was considered insufficient when the evaluation was carried out, but is highly likely to develop positively. A score of 4 is awarded when the impact is considered insufficient, and is highly unlikely to improve. In the final analysis, the scores 1 to 3 indicate that a project is 'sustainable', while 4 indicates 'unsustainable'. At the same time, sustainability carries a relatively strong weight in the overall assessment of a project. For example, a project can only be rated as 'performing well' overall (scores 1 to 3 out of 6) if it is also rated as 'performing well' by the criterion sustainability. Only the criteria 'effectiveness' and 'impact' carry a similar weight.

2.2
The conceptual framework of the meta-evaluation regarding the assessment of sustainability

Given the broad debate on the sustainability principle in development cooperation and the systematic performance rating of German development cooperation projects in relation to the DAC criteria, the present meta-evaluation proceeds on the assumption that sustainability has already been understood as a comprehensive and complex concept for some time. It also assumes that the understanding of sustainability is based on the two aforementioned aspects of sustainability, namely (i) sustainable development and (ii) the continuation of development results over time (see Sections 1.1 and 2.1). Logically, this kind of comprehensive underlying understanding of sustainability in evaluations only becomes evident in practice when all the DAC criteria are considered as a whole, as key aspects of sustainability only emerge as impact is substantiated. The conceptual framework for this empirical study of sustainability in the comprehensive sense therefore needs to include sustainability-related aspects from all five areas of performance assessment (i.e. the five OECD-DAC evaluation criteria). When we looked at the key questions for all DAC criteria synoptically, we identified a total of seven (distinct) areas specifically related to sustainability that are also frequently mentioned in the literature in conjunction with sustainability. These areas are described one by one below.

According to the BMZ guideline, the assessment of the OECD-DAC-based evaluation criterion 'sustainability' should take into account the stability of the context of a development project (BMZ, 2006). Analysis of **1) the context of a development project**, so the guideline recommends, should be based on the factors 'social justice', 'economic performance', 'political stability' and 'ecological balance'. Ultimately, analysing contextual factors will facilitate a sound examination of the external risks and potentials for the continuation of development results over time.

According to the logic of the DAC criteria, further criteria for assessing sustainable development performance drawn from the context of **2) the implementation of projects** are also important, such as participation by partners and target groups in implementation processes, and alignment with partner-country priorities. Such elements of the international aid effectiveness agenda are key components of the criteria relevance and effectiveness (BMZ, 2006).

Also important when analysing sustainable development are findings concerning **3) the outcomes** of a development project, i.e. the project's short- and medium-term results (Ashoff, 2015). In addition to the quantity and quality of projects, other important aspects include the changes they prompt, for instance with respect to ownership, awareness and resilience among local actors, and the reach which this entails (Boone, 1996). In the BMZ guideline, outcomes are discussed chiefly in conjunction with the criterion effectiveness, though in some cases also in conjunction with the criterion impact.

In conjunction with the criterion sustainability, with regard to the key questions on the risks and potential the guideline also

[2] *The other four DAC criteria are rated along a scale from 1 to 6 (with 1 as the highest and 6 as the lowest score). Since 2014 the GIZ has been rating sustainability along a six-point scale based on a points system: 'performing very well' (14 – 16 points), 'performing well' (12 – 13 points), 'satisfactory' (10 – 11 points), 'slightly unsatisfactory' (8 – 9 points), 'unsatisfactory' (6 – 7 points) and 'highly unsatisfactory' (4 – 5 points).*

recommends a focus on **4) local capacities**. This generates information on the extent to which local partners, executing agencies and target groups will succeed in continuing the activities, outputs and results without external support. As the direct effects of the activities and outputs generated by German development cooperation projects diminish over time, and ultimately come to an end when the support expires, local capacities then gain greater relative importance as time progresses (van Tulder and Pfisterer, 2008). So far, local capacities have been discussed largely in conjunction with the evaluation criterion 'sustainability'.

Regarding causal relationships, the contributions made by a project to **5) impact** are a further integral component of the understanding of sustainability. These include the positive and negative, and primary and secondary, long-term effects generated by a project either directly or indirectly, and either intentionally or unintentionally. The intended results are usually assessed by comparing the planned project results with those actually achieved in relation to formulated overarching objectives and global agendas (such as poverty alleviation). Unintended effects are also included in the assessment. In this context the BMZ guideline makes explicit reference to determining effects at the level of impact when assessing sustainability. According to the OECD-DAC, impact is an evaluation criterion in its own right.

A further key aspect of the understanding of sustainability in German development cooperation is **6) the predictability of the continuation of results** (Caspari, 2004; OECD, 1991; Stockmann and Gaebe, 1993; Stockmann and Silvestrini, 2012). According to the BMZ guideline, at this point evaluators should assess the extent to which the positive results of the development project will continue once the support has ended (BMZ, 2006). The predictability of the continuation of results is the key aspect of the OECD-DAC-based criterion sustainability.

Ultimately, an analysis of results (under impact) encompasses not only the sustainability dimensions of social justice, economic performance, political stability and ecological balance, but also an analysis of potential synergies and/or conflicts between the dimensions (BMZ, 2006). The assumption is that by including all dimensions, synergies – and therefore more sustainable results – will be achieved (OECD, 2016a). The sustainability debate addresses **7) interactions between the dimensions of sustainability**. These dimensions should therefore be included when evaluating development cooperation (Cutter, 2014; Dietz and Hanemaaijer, 2012; Islam and Clarke, 2005). Due to the importance of sustainability, interaction between the dimensions was also included in the key principles of the 2030 Agenda (UN, 2015).

2.3 Evaluation practices in German Financial and Technical Cooperation

The purpose of evaluating development projects is to assess the overall performance of development cooperation. Pursuant to the 'Guidelines for bilateral Financial and Technical cooperation with Germany's development cooperation partners', the implementing organisations carry out their own evaluations of a meaningful sample of completed and, if appropriate, ongoing development interventions. They do so 'on the basis of procedures laid down in consultation with the German government and based on OECD-DAC criteria and standards for independent evaluations' (BMZ, 2008).

In accordance with these instructions, at the module level official German development cooperation has a high overall level of coverage by evaluations. At the GIZ, over the last ten years virtually all projects (variously referred to as modules or phases) have been subjected to at least one evaluation. At the KfW, at least half of all projects in each sector are evaluated. When we collected the data for this meta-evaluation in October 2016, there were 1,081 completed evaluations that had assessed the sustainability of a total of 1,269 projects since 2006. Various types of evaluation were used in this context. Several types of evaluation are used during the course of projects, in some cases to manage and plan follow-on projects. Ex-post evaluations, on the other hand, analyse the performance of completed projects retrospectively after a certain interval.

To assess the performance of Financial Cooperation projects the KfW uses exclusively ex-post evaluations, which are usually conducted three to five years after completion of the project in question. Since 2006 the projects have been selected on the basis of a fixed sampling plan that each year incorporates 50 per cent of the 'evaluation-ready'[3] projects in each sector. These evaluations are managed by the independent evaluation unit of the KfW Development Bank (FC Evaluation Department). The evaluations are carried out by staff of the unit together with so-called delegates, i.e. staff from other sections of the company, and with external consultants. The KfW's ex-post evaluations usually follow a standardised procedure. Once an evaluation concept has been drawn up a questionnaire is sent to the project executing agency. The next step is to evaluate any available monitoring and final review reports. This is followed by an evaluation mission, which is sometimes supported by independent technical experts, and finally by preparation of an evaluation report. According to information supplied by the KfW the entire process of an ex-post evaluation takes around 37 working days, 27 of which are required for the evaluation itself and 10 for quality assurance by the evaluation unit.

Since 2006, GIZ has organised the evaluation of TC projects on both a centralised and a decentralised basis. The centrally organised types of GIZ evaluation include final and ex-post evaluations. The decentralised types are today's project evaluations (PEs) and the earlier project progress reviews (PPRs). The four types of GIZ evaluation that we looked at are described briefly below.

PEs are a type of evaluation that, when a follow-on project is planned, also includes the project appraisal. PEs were introduced in April 2014, and form today's GIZ evaluation format for modules. As a rule the latter are conducted twelve to six months before projects come to an end. A PE begins by defining the object of the evaluation and drawing up the evaluation design. This involves defining which activities will serve the purpose of evaluation, and which activities will serve the purpose of appraisal. This is followed by the collection of data (in the project setting) at a kick-off workshop together with the evaluation stakeholders. At a final workshop the provisional findings are presented based on the system of OECD-DAC criteria. Responsibility for accepting the evaluation report rests with the officer responsible for the project commission. Prior to that the report is subjected to quality control by the GIZ Evaluation Unit. Responsibility for accepting the published summary report rests with this unit. For PEs without a follow-on project an average of 49 working days are required. For PEs with follow-on projects the figure is 74 working days, though it is not clear how many of those days are used for the evaluation and how many for the appraisal to plan the follow-on project. The terms of reference can be handled flexibly, depending on whether it is a particularly complex project or whether the level of complexity is expected to be moderate or low. The Evaluation Unit requires an estimated one working day for quality assurance of the summary report.

The earlier PPRs – since superseded by PEs – always combined elements of evaluation and appraisal in a single format. There was no such thing as a PPR without a follow-on phase. The object of evaluation was the relevant phase of the development project. The PPR process was already similar to the PE process. PPRs were also usually conducted twelve to six months prior to the end of projects, and they too were preceded by a process of discussion with the partners in the project setting. As with PEs today, responsibility for managing a PPR rested with the officer responsible for the project commission. PPRs underwent quality control by the Evaluation Unit on a selected sample as part of GIZ meta-evaluations. There were no specifications for the total number of working days, though on average approximately 23 working days were required for preparation, implementation and analysis.

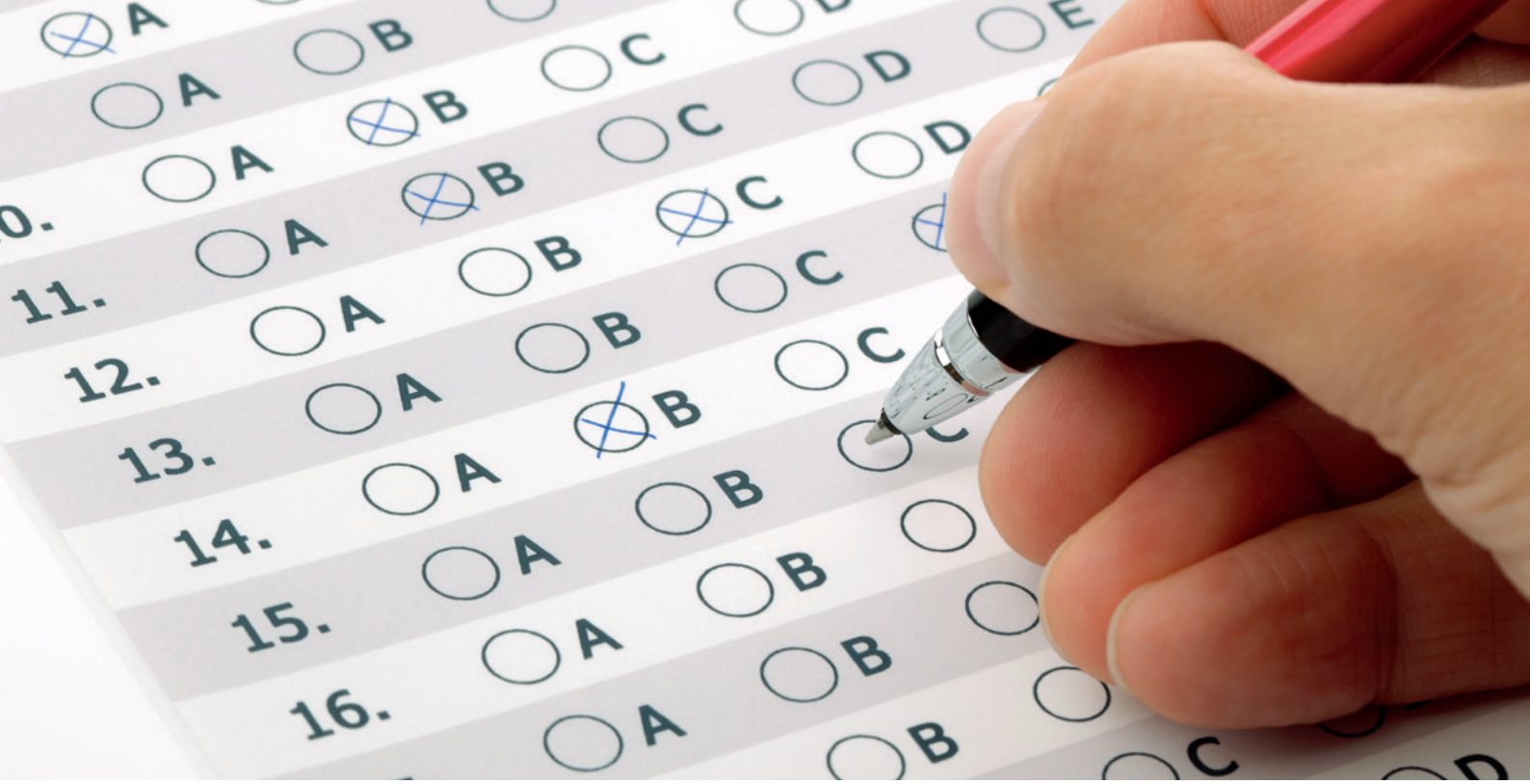
The Evaluation Unit was responsible for, and designed and managed, the GIZ's final and ex-post evaluations under the former independent evaluation programme[4]. This involved analysing individual sectors over a specific period of time. Independent institutions and consulting firms were usually commissioned to conduct these evaluations. Final evaluations were usually conducted between six months before and six months after the end of the project in question. Ex-post evaluations were held two to five years after the end of the

[3] At the KfW, projects are considered 'ready for evaluation' that were completed at least three years prior to sampling.
[4] The independent evaluations also included ex-ante and interim evaluations. Due to their low explanatory power for sustainability these types of evaluation were not included in this meta-evaluation.

project. For carrying out the evaluation 42 days were specified for the international consultant and 30 days for the national consultant; in individual cases where the evaluation methodology was particularly complex the figures could be higher. Approximately 12 working days were allowed for management and quality assurance by the unit. In contrast to PPRs and PEs, the object of these evaluations was the development project throughout its lifetime, including all phases. An inception report was also published.

Ultimately, however, the complexity of individual evaluation instruments can only be compared to a limited extent, as PPRs and independent evaluations involve proposed planned values, whereas the figures for PEs involve actual values (GIZ, 2016).

Clear differences become evident when we compare the various evaluation formats of the KfW and GIZ. The GIZ's former final and ex-post evaluations were relatively complex and costly to implement, for instance, and were managed by the in-house unit. The KfW's ex-post evaluations are smaller in scope, and are supported by a system in which staff members act as delegates. Here too, quality assurance is performed by the in-house evaluation unit at the KfW's head office. These three formats thus differ from the GIZ's decentralised evaluations (PEs and PPRs), responsibility for which rests with the officers responsible for the respective project commissions, and which so far have been subjected to quality control by the unit on the basis of a selected and partial sample only. The work required to perform the former PPRs tended to be less than that required for the other evaluation types.



3.

METHODOLOGY

3.1 Database

The database comprised the KfW and GIZ projects evaluated since approval of the BMZ guideline on consistent use of the DAC criteria in 2006. The population included all projects whose sustainability had been assessed independently of each other in decentralised or centralised evaluations.

When determining the population it was necessary to bear in mind that GIZ and KfW projects often comprise a chronological sequence of phases/modules involving continuity of content. While final and ex-post evaluations are not followed by a further phase/module of the project, when a PPR or a PE is carried out there may be a further phase/module of the project, and therefore a subsequent evaluation. To capture the latest possible assessment of sustainability, we included only the most recent evaluation of each project in the population.

When we collected the data in October 2016, 1,015 projects met these conditions. From the field of Financial Cooperation 462 ex-post evaluated KfW projects were included in the population. From among GIZ's centralised evaluations, 56 ex-post and 44 finally evaluated projects were included. From the decentralised evaluations 110 projects were included that had been subjected to PEs, along with 343 that had undergone PPRs (see Table 1).

For the purposes of the present meta-evaluation we analysed a representative sample of the population described. This took into account different types of evaluation and the distribution of sustainability scores. In formal terms, a randomised sample stratified by evaluation type was drawn that for each type was representative of both the mean value for the distribution of scores along the four-point scale (1 – 4), and the binary distinction between 'sustainable' (score 1 – 3) and 'unsustainable' (score 4) projects (see Table 1). A total of 513 projects were thus included in the sample.

This meta-evaluation is divided into two parts. The analysis of evaluation quality and the analysis of sustainability assessment are therefore presented below in two sections together with their respective methodologies.

Table 1: Overview of the database

	Type of evaluation	Timing relative to end of project	Number of evaluated projects	Number of evaluated projects in sample
GIZ	PPRs	12 to 6 months before end	343	174
	Final evaluations	± 6 months before/after end	44	38
	PEs	12 to 6 months before end	110	82
	Ex-post-evaluations	2 to 5 years after end	56	47
KfW	Ex-post-evaluations	3 to 5 years after end	462	172
Total			**1,015**	**513**

Source: Authors' own table

3.2 Evaluation quality

To facilitate a sound analysis of the assessment of sustainability based on evaluations, we first of all need to analyse the robustness of the evaluation findings. Here we are proceeding on the assumption that analysing the overall quality of an evaluation also permits us to draw conclusions concerning its specific quality regarding evaluation of the criterion sustainability. This analysis of evaluation quality forms the first part of this meta-evaluation.

A meta-evaluation is also referred to as an 'evaluation of evaluations' (Patton, 2008; Scriven, 1991, 2009). The purpose of a meta-evaluation is to systematically analyse the quality of evaluation processes and the robustness of the conclusions drawn (Leeuw and Cooksy, 2005). To allow ourselves to compare the quality of individual evaluations, we first of all need to define standardised criteria of the quality of the evaluation reports being studied.

When developing the evaluation grid (Table 2) for quality assessment we drew on findings from evaluation research (Patton, 2008; Scriven, 2009; Stufflebeam, 2001; Widmer, 2006) and examples of the way evaluation methods are applied in development cooperation (Carlsson and Wohlgemuth, 1996; Hageboeck et al., 2013; Leeuw and Cooksy, 2005). We also took into account the KfW's and GIZ's internal regulations for evaluation practice. Further guidance was also provided by the existing – though as yet unpublished – meta-evaluations in the field of German TC. After that we performed a pre-test on the evaluation grid using selected reports by evaluation type.

The final evaluation grid contains six areas of analysis with a total of 16 quality criteria. For each report in the sample, all criteria were rated as being either 'met' or 'not met'. For each assessment criterion included in the evaluation grid we produced a definition. For a detailed description of the assessment criteria used together with their definitions, please refer to Table 4 in the Annex. Throughout, the basis on which we analysed evaluation quality was the report in its entirety, i.e. all written documents of the evaluation including annexes. To analyse the quality of reports, we first of all fed our criteria grid and the reports to be analysed as PDF files into the qualitative analysis programme 'MAXQDA' (a software application). The next step was to store in a database our judgement, based on our reading of the reports, of whether a criterion was met, using the data management programme 'Microsoft Access'. We used the software to reference the point in each report on which our judgement was based, so that we would then be able to reconstruct our assessments.

To test the intersubjective comparability of these judgements within the evaluation team, 10 per cent of the analysed reports – stratified by evaluation type – were encoded several times, i.e. read and assessed by different people. We then used Cohen's Kappa intercoder reliability coefficient to determine the degree of inter-evaluator consistency of encoding behaviour. The Kappa value for quality assessment is 0.62, which indicates substantial agreement (Landis and Koch, 1977).[5]

In addition to a descriptive analysis of the quality criteria, an aggregate 'quality index' also allows direct comparison between evaluation reports. To form the index we first of all added up the number of criteria met. Since the focus of our analysis was on the quality of conclusions concerning the assessment of project sustainability, criteria that supply information on the robustness of the findings (Q-9 to Q-16) were weighted double. This meant that one evaluation could achieve a maximum of 24 points. Finally, to facilitate interpretation we divided the value achieved in each case by the maximum number of 24 points, to obtain an index with values along a scale of 0 to 1.

[5] *The standard or most frequently-cited interpretation of the Cohen-Kappa coefficient dates back to a study by Landis and Koch (1977), who propose the following scale of interpretation: '0.01 – 0.20 = slight agreement, 0.21 – 0.40 = fair agreement, 0.41 – 0.60 = moderate agreement, 0.61 – 0.80 = substantial agreement, 0.81 – 0.99 = almost perfect agreement'.*

Table 2: Overview of quality criteria

Areas	Criteria
1) Evaluation background	Q-01 Object described
	Q-02 Area of enquiry formulated
2) Explication of the causal relationships	Q-03 Results logic described
	Q-04 Indicators formulated
3) Methodology	Q-05 Methodology described
	Q-06 Strengths and limitations of the evaluation discussed
	Q-07 Stakeholder respondents identified
	Q-08 Selection procedure described
4) Evaluation design	Q-09 Before and after comparison
	Q-10 Control / comparison groups
	Q-11 Causality inferred on the basis of plausibility
5) Robustness of the findings	Q-12 Triangulation of data
	Q-13 Triangulation of methods
6) Analysis/conclusions	Q-14 Conclusions referenced
	Q-15 Conclusions plausible
	Q-16 Database adequate

Source: Authors' own table

3.3 Sustainability assessment

The second part of the meta-evaluation involves the analysis of sustainability assessment criteria. In harmony with the logic of the quality analysis (described in Section 3.2), we entered individual assessment criteria in an assessment grid which then formed the framework for the quantitative content analysis (see Table 3). We then extended the traditional design of a meta-evaluation as a quality analysis to include the analysis of the specific assessment criteria. Ultimately, it is only this thematic extension of the meta-evaluation that allows us to analyse comprehensively the evaluation and assessment of sustainability in German development cooperation.

The conceptual framework for the evaluation grid to analyse the sustainability assessment criteria was provided by the BMZ guideline on applying the OECD-DAC evaluation criteria (see Section 2.2). The areas we drew from it guided us in identifying specific criteria that, as expected, are used to assess project sustainability. Furthermore, by analysing the guidelines of the KfW and GIZ we also collected theoretically possible criteria. We also compared our approach with the current literature on the evaluation of sustainability. However, the high conceptual complexity of sustainability leads us to assume that a purely deductive approach would be unable to fully capture the underlying conceptual understanding of sustainability that practitioners have. We therefore supplemented the deductive approach with an exploratory study of 40 KfW and GIZ evaluations. The study was designed to compare theory and practice, taking into account specific features of FC and TC projects, of different types of evaluation,

and of evaluation and assessment practices across time. Once again we tested the analysis grid using selected reports.

The findings we generated are shown in Table 3. The criteria for assessing sustainability are broken down according to the areas described in Section 2.2., namely: 1) context, 2) implementation, 3) outcome, 4) local capacities, 5) impact (unintended effects,) 6) continuation of results, and 7) interaction between the dimensions of sustainability. These areas form the conceptual framework for 18 sustainability criteria, which we broke down further by actor, sustainability dimension and capacity type into 48 criteria. We also included the overarching and programme objectives in our analysis, and assigned them to the dimensions of sustainability and the SDGs.

Table 3: Overview of sustainability criteria

Areas	Criteria	Differentiated criteria
1) Context	1. Context by dimension	S-01 Social dimension
		S-02 Economic dimension
		S-03 Political dimension
		S-04 Environmental dimension
2) Implementation	2. Alignment	S-05 Alignment with national rules
		S-06 Alignment with the sociocultural context at the level of target groups
	3. Participation	S-07 Participation by the development partner
		S-08 Participation by target group(s) / population
	4. Management	S-09 Use of local (institutional) structures
		S-10 Management response / learning from M&E / lessons learned
		S-11 Scaling-up implemented
		S-12 Exit strategy in place
3) Outcome	5. Acceptance and ownership	S-13 Acceptance and ownership by the private-sector agency
		S-14 Acceptance and ownership by the partner
		S-15 Acceptance and ownership by the target group
	6. Outputs of the executing agency/partner	S-16 Service / product quality
		S-17 Service / product quantity
	7. Use of outputs	S-18 Use of outputs by the partner / executing agency
		S-19 Use of outputs by the target group
	8. Change of awareness	S-20 Change of awareness in the partner / executing agency
		S-21 Change of awareness in the target group
	9. Resilience and adaptability	S-22 Resilience and adaptability of the partner / executing agency
		S-23 Resilience and adaptability of the target group
	10. Reach	S-24 Structure-building
		S-25 Dissemination

Areas	Criteria	Differentiated criteria
4) Local capacities	11. Capacities of the partner	S-26 Financial capacities
		S-27 Human capacities
		S-28 Institutional capacities
	12. Capacities of the executing agency	S-29 Financial capacities
		S-30 Human capacities
		S-31 Institutional capacities
	13. Capacities of the target group	S-32 Financial capacities
		S-33 Human capacities
		S-34 Institutional capacities
5) Impact[6]	14. Unintended effects by dimension	S-35 Social dimension
		S-36 Economic dimension
		S-37 Political dimension
		S-38 Environmental dimension
6) Predictability of the continuation of results	15. Predictability of the continuation of results by dimension	S-39 Social dimension
		S-40 Economic dimension
		S-41 Political dimension
		S-42 Environmental dimension
7) Interaction between the dimensions of sustainability	16. Synergy between the dimensions	S-43 Creation of synergies by projects
		S-44 Identification of synergies by the evaluation
	17. Conflict between the dimensions	S-45 Identification of conflicting objectives by the project
		S-46 Identification of conflicting objectives by the evaluation
	18. Side effects tolerable	S-47 Classification of possible compensation measures by the project as sufficient and / or of possible side-effects as 'tolerable'
		S-48 Classification of possible side effects by the evaluation as 'tolerable'

Source: authors' own table

We included in the quantitative content analysis only those criteria which, according to the evaluation report, were directly related to sustainability.[7] In all cases we performed the analysis on the basis of the report in its entirety, i.e. all written documents of the evaluation including annexes. Regarding the sustainability-related conclusions we subsequently tested whether the evaluation report indicated that a criterion was either present or not present (e.g., whether ownership existed or not). Where a report did not give any clear indication regarding a criterion, we defined the presence or absence of a criterion as 'unclear'. Furthermore, we recorded in the data survey whether the presence or absence of the criteria in

[6] *The area 'impact' includes both 'intended' results and 'unintended' effects (see Section 18). However, since the 'intended results' are an integral part of the assessment of the OECD-DAC criterion 'impact', we included in the sustainability assessment grid only the criteria for 'unintended' effects. We recorded the 'intended results' separately. The findings are presented in Section 47 .*

[7] *This means that we only used as a basis for our assessment those points in the reports where the criterion in question was linked 1) with the word 'sustainability', 2) with impact, 3) with its continuation over time, 4) with a risk assessment or 5) with interaction between the dimensions of sustainability.*

question had an enabling or constraining effect on sustainability, or whether its effects were unclear.[8]

When testing the intercoder reliability of sustainability assessment we obtained a Kappa value of 0.63. Thus the overall Kappa value for quality and sustainability assessment[9] was 0.63, indicating substantial agreement (Landis and Koch, 1977).

3.4 Contextual study

The methodology described so far enables us to systematically analyse evaluation and assessment practices in German development cooperation. Whether or not these practices are also appropriate can only be determined by international comparison, however (see Evaluation Question 3). In this meta-evaluation we perform this comparison in the form of a contextual study devoted to evaluation and assessment practices of other bi-and multilateral development organisations. Since the DAC criteria of 1991 form the basis of sustainability assessment for evaluation units in the OECD countries (OECD, 1991), in the contextual study we investigated how these units apply sustainability as an assessment criterion. We also included in the analysis selected multilateral organisations with sophisticated approaches to evaluating sustainability.

The population for the study comprises 40 evaluation units from 37 member countries of the OECD-DAC Network on Development Evaluation (EvalNet), plus nine multilateral organisations whose evaluation systems were analysed in detail in the current round of the DAC Peer Review process (OECD, 2016b).[10]

The database for the contextual study was provided by the evaluation units' guidelines on applying the evaluation criterion that are available online. Here we selected a step-by-step procedure. First of all we screened the websites to see if they provided a transparent description of the units' actual evaluation practices. We then included 24 evaluation units on whose assessment system sufficient information was available in the comparative analysis of assessment systems. These included 18 bi- and six multilateral evaluation units of the countries Australia, Austria, Belgium, Canada, Denmark, France, Japan, Luxembourg, the Netherlands, New Zealand, Norway, Sweden, Switzerland, the United Kingdom, the United States and Germany, as well as the African Development Bank, the Asian Development Bank, the European Investment Bank, the European Commission, the United Nations Development Programme and the World Bank Group. Our comparison focused on both the definition of the criterion of sustainability, and the assessment practices described. It emerged that particularly in Switzerland and the USA, and in evaluations performed by the World Bank and the African Development Bank, the underlying concept of sustainability was a comprehensive one. At least 8 out of 39 criteria were used to evaluate sustainability. The key criteria are financial, political, technical and social sustainability, plus ownership. Finally we performed an in-depth analysis of sustainability assessment by three bi- and multilateral evaluation units that in addition to individual assessment criteria also use rating scales (involving the award of scores or points), and in this respect are highly comparable in their evaluation practices to German development cooperation.[11]

[8] *When developing the grid of criteria we proceeded in a similar way as in the quality analysis. We began by feeding all the evaluation reports and sustainability criteria to be studied into MAXQDA, and then encoded them using a Microsoft Access database. Here too we tested the intercoder reliability by double coding 10 per cent of the evaluation reports, stratified by evaluation type (see Section 3.2).*

[9] *The overall assessment is based on the aggregate analysis of agreement in the assessment of quality and sustainability performed by three evaluators. The overall Kappa value is the mean of the values for quality and sustainability criteria. A value of 0 indicates maximum divergences between the evaluators; a value of 1 indicates maximum agreement between the evaluators (see Section 30 for further explanation).*

[10] *A first study on the status of evaluation systems was performed in 2010 (OECD, 2010a).*

[11] *The steps 'screening the websites of the evaluation units' and 'comparative analysis of 24 evaluation units' were performed by DEval. The in-depth study was conducted in cooperation with Jana Preiß as part of her master's thesis at the Freie Universität Berlin (Preiß, 2017).*

3.5 Limitations

This meta-evaluation is a desk study based on secondary data. The depth of analysis was therefore determined by the reports drawn up in accordance with the GIZ and KfW guidelines for the respective types of report. Since sustainability forms only a part of an evaluation of project performance, and the relevant comments contained in the report were correspondingly succinct, it was not always possible to identify a criterion as an enabling or constraining factor for project sustainability. This meant that in many cases we had to encode the positive or negative effect of a criterion as 'unclear', and exclude it from the analysis as a result. Having said that, the number of points in the reports encoded as 'unclear' did not exceed the cases identified as 'clearly positive or negative' for any criterion, hence we may assume that this fact is of little significance. However, it may indeed carry some weight for criteria on which the reports had little to say.

The level of detail in reporting also plays a role in the assessment of reporting quality. This was based solely on the assessment of the evaluation reports. The different instructions regarding the degree of detail when reporting possibly leads to discrepancies between the actual quality of an evaluation and the quality that can be discerned on the basis of the evaluation reports. To minimise these discrepancies, when selecting the evaluation criteria we were careful to include only those criteria in the quality assessment that theoretically would have to be met by a large number of evaluation types because they are of more fundamental importance with regard to quality.

When analysing the understanding of sustainability in relation to the criteria included in the reports, we ultimately came up against a number of challenges with regard to endogeneity. It is possible that certain sustainability criteria are discussed more frequently or in greater detail in evaluations because they have a particularly positive or negative effect, whereas neutral effects tend to be emphasised less frequently. Another possibility is that negative/positive manifestations of a criterion are easier/more difficult to demonstrate methodologically. Furthermore, the different guidelines and expectations associated with a specific type of evaluation may also influence evaluation findings. Our discussion of findings (Chapter 4) therefore takes account of systematic differences in the aforementioned respects, and where necessary draws attention to possible limitations to the explanatory power of the findings.

A further limitation in quantitative content analyses is the intersubjective comparability of coding behaviour between two or more evaluators. There is a risk that different individuals may interpret one and the same fact differently, and reach different findings as a result. We therefore tested the intercoder reliability of the evaluation team using the Kappa value after Cohen (see Sections 3.2 and 3.3). The overall Kappa value of 0.63 demonstrates moderate or substantial agreement between the evaluators when assessing quality and sustainability. Since this value can be interpreted as demonstrating strong agreement, though not very strong agreement, we obtained an additional external perspective on quality assessment. We compared the quality assessments of reports analysed in this meta-evaluation with those analysed in the GIZ meta-evaluations. We found that the assessments of what percentage of the maximum number of points was achieved were similar. Compared to the GIZ meta-evaluation in the health sector (Raetzell and Krämer, 2013), in the majority of projects analysed the assessment differs by less than ten per cent (where 100 per cent indicates that all criteria are met), and in only one case does the discrepancy exceed 20 per cent. Since no meta-evaluations of KfW evaluations are available as yet, it was only possible to perform this kind of comparison for GIZ evaluations.

4.

FINDINGS

In this Chapter we present the findings of the meta-evaluation in detail. Section 4.1 is devoted to the findings on evaluation quality. In Section 4.2 we discuss our findings on the understanding of sustainability in relation to the assessment criteria. Finally, Section 4.3 presents the findings from the contextual study.

4.1 Quality of the evaluation reports

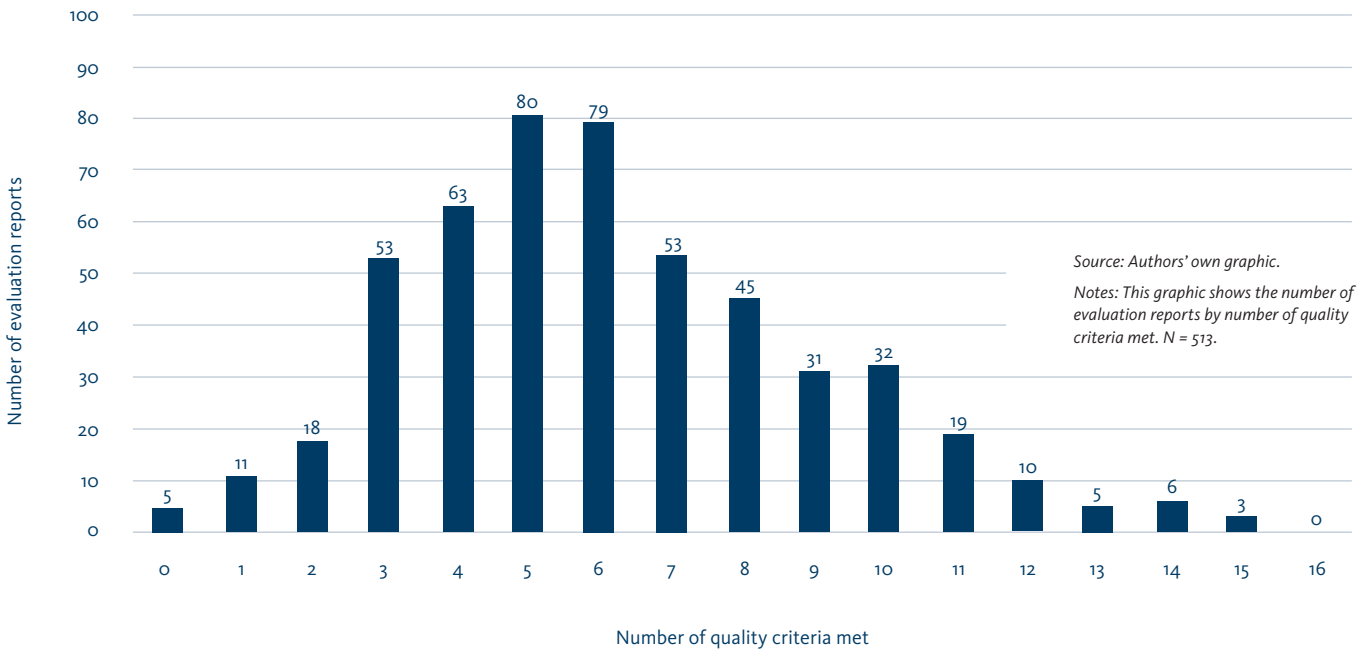
Analysing the quality of evaluation reports in the meta-evaluation enabled us to assess the robustness of the evaluation findings regarding sustainability. Recalling what we said about the methodological limitations (see Section 3.5), we will discuss the findings with caution, since we are unable to entirely rule out the possibility of discrepancies between the actual quality of an evaluation and the quality that we are able to identify from the evaluation reports. The findings on methodological quality are thus always to be seen in relation to the transparency of quality based on the existing evaluation reports. We only analysed additional documents such as inception reports or terms of reference when they formed part of the annexes of the evaluation reports. Contextual information, for instance on the resources used for the evaluation in question, were only rarely available and were therefore included in the analysis only at a general level. However, the findings of the meta-evaluation do permit us to conclude that the procedures we selected were entirely appropriate regarding the comparability of the individual evaluation reports. The normal distribution of the number of quality criteria across all reports (Figure 1) demonstrate that the reports analysed cover the entire spectrum of the quality grid we drew up. On average 6.2 of the 16 possible quality criteria were met.

The analysis of the quality areas (Figure 2) shows that the vast majority of the evaluation reports clearly describe the background of an evaluation (93%), the causal relationships (85%) and the methodology (84%). A quality area was considered as having been covered if at least one of the relevant quality criteria was met. A much lower percentage of evaluation reports addressed the evaluation design (25%) and the robustness of findings (33%). The findings of this aggregate analysis provide a first impression of the areas in which the evaluations did particularly well or not so well.

The findings on the quality of the evaluation reports show that virtually all evaluations (92%) describe their object (Q-01) (see Figure 4). However, ultimately this also means that not all evaluations provide sufficient information to show readers precisely what the evaluation is about. A low transparency of information becomes apparent particularly with respect to the criterion of operationalisation of the area of interest (Q-02) with respect to the standardised key questions based on the OECD-DAC criteria. Only in 16 per cent of cases is an object-specific area of enquiry evident from the evaluation reports, i.e. only in these cases are evaluation questions relating to the DAC criteria included that are geared to the specific object (Q-02). A supplementary analysis of selected additional documents of the KfW and GIZ shows that although the area of enquiry for an evaluation can be reconstructed from additional documents – such as the concept paper or the terms of reference – it is not evident from the actual evaluation report alone. We might therefore assume that the implementing organisations do not see their evaluation reports as stand-alone products that can be understood without additional documents.

In the majority of KfW and GIZ evaluations results are substantiated by comparing actual values with target values for selected indicators that form part of the results logic. The findings of this meta-evaluation indicate that the preconditions for proceeding in this way were created in most evaluations. The majority of reports presented the results logic (Q-03, 63%) and the corresponding results indicators (Q-04, 74%). In approximately one third of the projects we analysed the results logic was not made transparent in the evaluation reports, though this does not exclude the possibility that such logics were used. On the other hand, the presence of a results logic and results indicators is not a sufficient condition for drawing causal conclusions based on comparisons of actual values with target values. In only few cases do the reports respond to the challenge of causal attribution by incorporating more complex procedures for results analysis. Only 19 per cent of the evaluations included before and after comparisons (Q-09). One possible reason for this might be that barely any

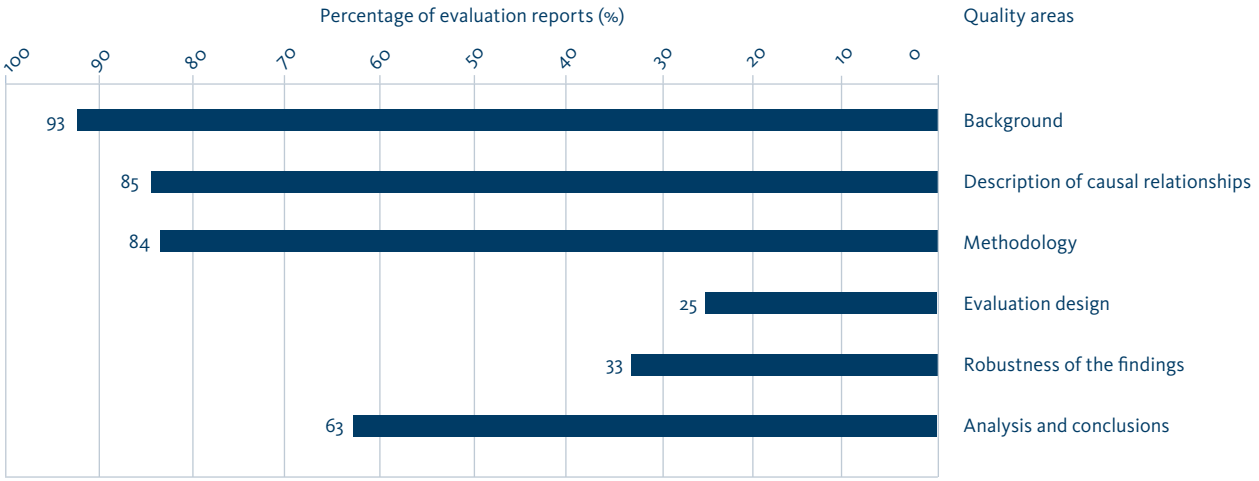
Figure 1: Number of evaluation reports by number of quality criteria met



baseline data are available for projects pursuing particularly innovative approaches or establishing fresh points of departure. In such cases baseline data can only be reconstructed using secondary data, which are then very difficult to compare with the current status when performing evaluations. Only 9 per cent operate with control groups. More complex theory-based procedures such as contribution analysis, which address the problem of attribution by applying systematic methods for plausibly associating possible causes with possible effects, have barely been used to date. This finding is also important with regard to sustainability, as the substantiation of results forms the key basis for assessing sustainability. We therefore also need to examine whether and to what extent the quality of an evaluation also has an empirical effect on the assessment of sustainability. These findings are presented later on in Section 4.3.

Working with baseline data, control groups or systematic methods for plausibly associating causes with effects is absolutely essential for robust results analysis. Without such methods, causal attribution is not permissible. In these cases, this uncertainty regarding causal relationships can only be reduced by using systematic triangulation methods. Around one third of the evaluations used systematic data triangulation procedures. Sufficient evidence that different methods were compared was provided in only just under one in ten cases. In this connection it is astonishing to note that when comparing actual values with target values evaluations only rarely make transparent use of monitoring data. In only 31 per cent of evaluations was information from the monitoring systems of the implementing agencies and/or partners and executing agencies explicitly included in the analysis strategy (see Figure 4). This does not mean that the evaluations had access to monitoring data only in approximately one third of cases. On the contrary: One may assume that the consultants are always provided with monitoring data by the projects and the executing agencies. The fact that barely any reference is made to these data in the conclusions drawn by evaluations rather points to the fact that they often do not match the purpose or the requirements of an evaluation. This also explains why some of the evaluations we analysed also indicate that the projects should in future invest more in establishing and

Figure 2: Percentage of evaluation reports by quality areas covered



Source: Authors' own graphic.

Notes: The graphic shows the percentage of evaluation reports that meet at least one criterion for each of the respective areas. N = 513.

maintaining results-based monitoring systems.

The findings of the meta-evaluation demonstrate that the substantiation of results could be made more robust by using relevant methods for causal analysis and triangulation. At present, only around one third of the evaluations indicate that their conclusions are founded on a sufficient database (Q-16). This assessment is based on the information on evaluation design and data collection methods. One reason for this is certainly the low availability of data, particularly in cases where projects in fragile contexts are being evaluated. The weaknesses of a data source could be reduced through systematic triangulation methods, however; at present only approximately 30 per cent of evaluations are using them. Furthermore, conclusions that are not founded on a robust database, but are necessary for the purpose of the evaluation, could be identified as such by explaining clearly the remaining uncertainty.

Moreover, the majority of findings and conclusions are substantiated plausibly (Q-15), but seldom referenced (Q-14).

Evaluation quality could be raised through two measures, namely: improving methodology in order to provide better substantiation of results and sustainability, and making the evaluation findings more transparent.

With regard to transparency, we also found that the methodology was described in only 68 per cent of cases (Q-05). In these cases we reconstructed the methodology on the basis of the report. The majority of evaluations included field missions, and used various data collection methods in the field. Eighty-three per cent of evaluations used semi-structured interviews, 26 per cent used group discussions and 13 per cent used standardised surveys to gather data (see Figure 3). Fifty-eight per cent of the reports described the surveyed groups. However, only 15 per cent of the evaluation reports describe the selection procedure used; in the remaining cases the selection appears arbitrary from the reader's point of view.

A synopsis of all criteria reveals that none of the 513 evaluation reports meets all 16 quality criteria (see Figure 1). However,

evaluations that do not meet all quality standards can certainly generate credible findings. With regard to causal analysis, for instance, it is not always necessary to combine before and after comparisons with control group comparisons and additional theory-based designs, even if this does always make the substantiation of results more robust. Generally speaking, the appropriate design for any particular evaluation is determined by the question it sets out to address, and the attributes of its object. As shown above, however, these questions are rarely evaluation-specific; they usually arise from the guidelines. Consequently, selection of the appropriate design for the GIZ's and KfW's module evaluations is based chiefly on the characteristics of the object of the evaluation, and on which designs are available and feasible. Additionally, however, the evaluations also include methods of analysis that are specific to the respective implementing organisations. Due to their lack of comparability, however, we did not include them in this meta-evaluation. Evaluations of FC projects in the economic infrastructure sector, for instance include micro- and macro-economic calculations that play no part in the evaluation of TC projects.

For comparative analysis of quality we use the quality index described in Section 3.2. With this index, an evaluation report that meets all 16 quality criteria is assigned the value 1. As described in Section 3.2, criteria that are particularly important for assessing the robustness of findings are weighted double. A report that does not meet a single criterion is assigned a value of 0. Across all 513 evaluation reports we studied, an average quality index value of 0.34 was achieved. This finding shows that a high number of the reports appear not to meet all the quality criteria that we applied in a transparent manner. Particularly regarding the criteria for robustness of the substantiation of results and sustainability, there is potential for raising methodological quality.

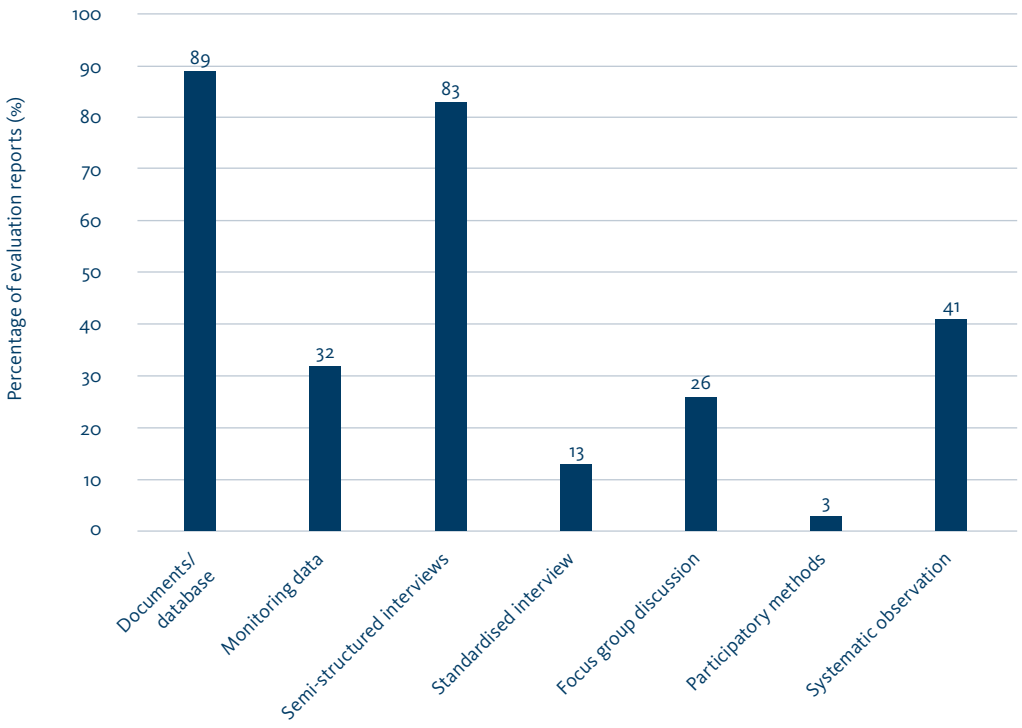
When we disaggregated our analysis of quality by evaluation type (see Figure 5 below) our findings were as follows. The GIZ's ex-post and final evaluations display the highest quality, with a mean index value of 0.6. These are followed by the KfW's ex-post evaluations and the GIZ's PEs, with a value of approximately 0.3. The lowest quality was shown by PPRs, with an average index value of 0.2. The differences between these three groups are statistically significant.[12] These findings demonstrate that more extensive and sophisticated evaluations pay off. The GIZ's earlier ex-post and final evaluations were usually more extensive and complex. However, between 2006 and the end of the independent evaluation programme in 2014, only 100 such evaluations were conducted. By contrast, the KfW's ex-post evaluations and the decentralised evaluations are less elaborate, yet cover large sections of the portfolio of GIZ and KfW projects. Hence quality sits somewhat uncomfortably between the scope of an evaluation and the overall degree of coverage by evaluations. Overall, we note that the quality of the evaluation reports improves over time. Whereas evaluations conducted in 2006 achieved an index value of approximately 0.3, ten years later this value was just under 0.4. An analysis disaggregated by type of evaluation corroborates this. Over the period of analysis, particularly GIZ ex-post evaluations and final evaluations display a sharp increase in quality, which is less pronounced in GIZ PPRs and KfW ex-post evaluations. In the case of GIZ PEs only a slight change is observed, due to the short period of time involved.

4.2 The assessment of sustainability in GIZ and KfW evaluations

This section deals with the understanding of sustainability in German development cooperation. The analysis is based on the findings of the quantitative content analysis in relation to the sustainability assessment criteria. We structure our discussion of assessment on the conceptual framework for sustainability (see Section 2.2) : After several general findings (Section 4.2.1) we will discuss our findings in the areas of context (Section 4.2.2), implementation (Section 4.2.3), outcome (Section 4.2.4), local capacities (Section 4.2.5), impact (Section 4.2.6), predictability of the continuation of results (Section 4.2.7), and interaction between the dimensions of sustainability (Section 4.2.8). The underlying understanding of each of the sustainability criteria is shown in the overview of Table 5 in the Annex. This discussion of assessment also incorporates the findings of the quality

[12] *A Welch test shows that the groups differ significantly ($p < 0.01$), and that the differences between the types of evaluation are thus very probably not due to chance. A Games-Howell test to directly compare these groups corroborates this finding.*

Figure 3: Percentage of evaluation reports by data collection methods used



Source: Authors' own graphic.

Notes: The graphic shows the percentage of evaluations in which each of the data collection methods was used. N = 513.

analysis (from Section 4.1), in order to contextualise the findings.

As explained in Section 3.3, we developed the evaluation criteria on the basis of an integrated approach that includes both a deductive and an inductive component. When defining the individual assessment criteria we were careful to draw distinctions between them that were as clear as possible (see Table 5). Having said that, differences in the subjective perspectives of evaluators mean that a lack of absolute conceptual distinction between the criteria can never be entirely ruled out. From an empirical point of view, however, this risk would appear modest. Based on the correlations between individual criteria, only few links are statistically significant. Only the criteria 'acceptance and ownership by the target group', 'use of outputs' by partners and 'synergy between the dimensions of sustainability' displayed strong links with other sustainability criteria.[13] Overall, however, a separate discussion of the individual assessment criteria would appear permissible.

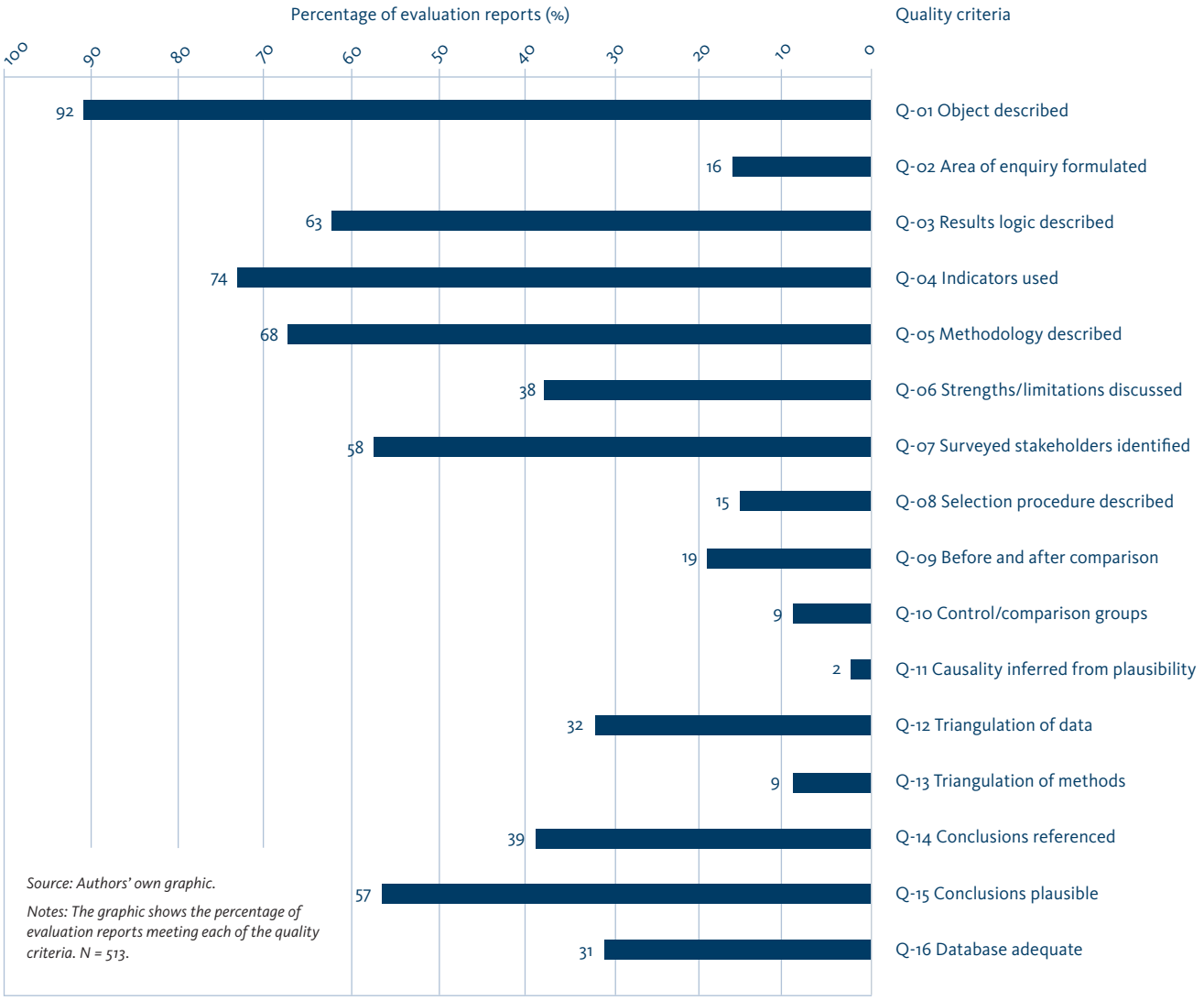
4.2.1 Overarching findings

Our synopsis of sustainability assessment practices clearly shows that these are based on a comprehensive understanding of sustainability. This is reflected in the large number of areas considered when performing an assessment. Figure 6 shows the frequency of the overarching criteria and areas (as a percentage) when at least one criterion from the area is included in the assessment. The overarching criteria are in turn based on more specific evaluation criteria.[14] Despite this broad base, there are areas which are included in the assessment significantly less frequently than others. This provides us with

[13] *Using Pearson's product-moment correlation coefficient, we analysed positive correlations from a strength of 0.7 and negative correlations from a strength of 0.5. We also analysed only pairs of variables that were included in at least 10 evaluation reports and that were significant at the 5 per cent level.*

[14] *We classified an overarching criterion as being 'included in the report' when one evaluation report made a positive or negative statement on at least one corresponding individual criterion with respect to sustainability. In the analysis below, however, we did not include in the data we collected any statements that were equivocal (i.e. neither positive nor negative).*

Figure 4: Percentage of evaluation reports by quality criteria met



Source: Authors' own graphic.

Notes: The graphic shows the percentage of evaluation reports meeting each of the quality criteria. N = 513.

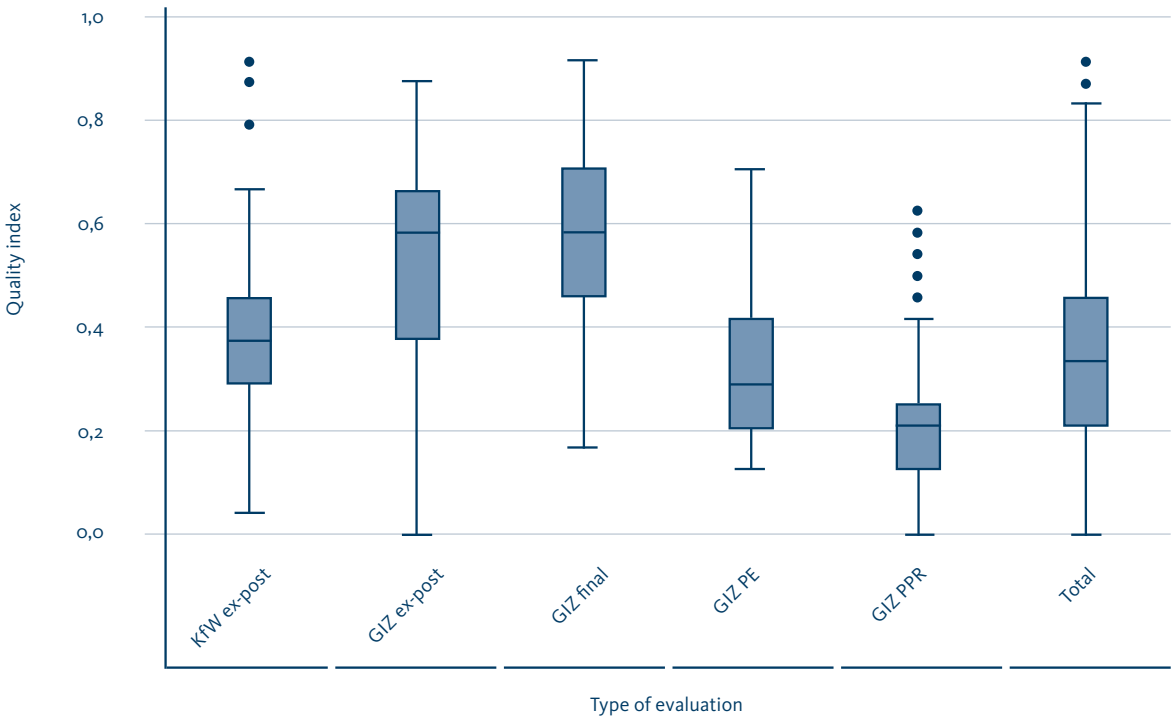
a first indication that assessment is not standardised across evaluations. Presumably, one reason for the low level of standardisation is insufficient guidance on how sustainability is to be understood conceptually in relation to sustainable development across the five DAC criteria.

Areas that are included in the discussion of sustainability in evaluations relatively frequently are 'outcome' (mentioned in 87 per cent of all evaluations) and 'local capacities' (mentioned in 86 per cent of all evaluations, see Figure 6).[15] This shows that direct effects and local capacities play an important role in assessment, and are therefore an integral component of the underlying understanding of sustainability.

The findings also show that the key questions contained in the BMZ guideline are certainly used, albeit not as frequently or systematically as expected. With regard to the evaluation criterion 'sustainability', the first two key questions require the

[15] *We classified an area as 'included in the report' when at least one of the criteria assigned to it was included in the assessment of sustainability. For the area 'local capacities' there are three possible criteria, whereas for the area 'outcome' there are five. This means that the area 'outcome' is likely to be classified as 'included in the report' more quickly.*

Figure 5: Quality index by type of evaluation



Source: Authors' own graphic.

Notes: The graphic shows the quality index by evaluation type. The quality index is formed using the quality criteria Q-01 to Q-16, shown in Figure 4. The quality criteria Q-9 to Q-16 receive are weighted double. A quality index value of 1 indicates the highest methodological quality of a report, and a value of 0 the lowest. The graphic shows the distribution of the data for each evaluation type as box plots. The boxes in the centre represent the middle 50 per cent of a distribution, bisected by the median. The horizontal lines above the boxes demarcate the values that are above the third quartile; the lines below the boxes demarcate the values that are below the second quartile. The dots represent the outliers. N = 513.

evaluator to examine the predictability of the continuation of results and the context of a project. The empirical findings of this meta-evaluation confirm that the relevance of these two areas is relatively high. However, contextual factors and the predictability of the continuation of results are included in the assessment of sustainability only in about one in two reports. The third key question concerning the evaluation criterion sustainability concerns the risks and potential in the project context. Empirically, the answer to this question is found in various criteria in the areas 'implementation' and 'outcome'.

However, it is also evident that a discussion of possible 'unintended effects' is included significantly less frequently in the assessment of sustainability. Since identifying unintended effects is one of the fundamental difficulties faced in evaluations, this finding is hardly astonishing. At the same time, however, it also shows that the evaluations fall short of their own aspirations at this point, since the discussion of unintended effects is from a conceptual perspective an elementary component of the criterion of impact. Furthermore, the BMZ's definition of the criterion impact suggests that evaluators should consider various dimensions of effects and relate these to each other where possible. This too occurs very rarely at present. Once again, the methodological difficulty of systematically identifying interactions between individual dimensions of objectives no doubt comes into play here. When analysing the quality of evaluation, the evaluations we looked at performed relatively poorly, especially with regard to the substantiation of results. Presumably, one reason for this is the lack of enabling

preconditions for evaluating unintended effects and interactions between the dimensions. To date, both aspects have rarely been an explicit component of the results logic of GIZ and KfW projects. However, the early identification of potential side effects of projects and interactions between projects is key to monitoring them later on.

The findings demonstrate that the existing understanding of sustainability clearly goes beyond the aspect of the continuation of development results over time, but does not yet coincide with the understanding of sustainability inherent in the 2030 Agenda.

When judging the understanding of sustainability based on the assessment criteria that we applied, however, we were interested not only in whether certain criteria were used to assess sustainability, but also in whether, in the opinion of the evaluators, the presence or absence of these criteria was considered an enabling or constraining factor for project sustainability. Figure 7 shows that according to the reports all the overarching assessment criteria can affect the assessment either positively or negatively. Here is an example: If an evaluation ascertains that acceptance and ownership are present amongst the partners, evaluators usually see this as a positive factor when assessing sustainability. On the other hand, if an evaluation determines that there is no partner ownership, evaluators see this finding as a challenge when assessing sustainability and ultimately correct the sustainability score downwards. Since the theoretical case in which a criterion, although present, has a negative effect on the assessment of sustainability, occurred only rarely, we did not include these cases in the analysis on a differentiated basis. An example of such a theoretical case would be the criterion 'use of outputs'. While the use of outputs is usually seen as a positive factor in the assessment of sustainability, the overuse of outputs would be seen in a negative light.

From the perspective of project management and evaluation, the question arises of which criteria are seen as having a largely positive or negative effect on sustainability. Here it emerged that the evaluations ascribe a positive effect to most of the criteria. Only the criteria 'context' and 'partner and executing agency capacities' are seen in most cases as constraining factors for sustainability. This finding is also transferred to the relevant areas. While 'context' and 'local capacities' are evidently seen as constraining sustainability, the areas 'implementation', 'outcome' and 'predictability of the continuation of results' do relatively well. The areas 'impact' and 'interaction between the dimensions' as seen as having a largely positive effect on sustainability; these areas were, however, included in the reports significantly less frequently.[16]

These findings reveal a tendency. While external factors in the context of the development project tend to be seen as constraining factors when assessing sustainability, criteria that lie within the sphere of influence of projects are more likely to be seen as conducive to sustainability. We will discuss this fact in further detail as we describe our findings on the individual sustainability areas below. It is also noteworthy that the GIZ rates the sustainability of its projects significantly more positively in its evaluations than the KfW. This is also corroborated by the accompanying evaluation synthesis (Noltze et al., 2018). When we compare findings across the regions, it is striking that in all areas except 'context', sustainability is assessed significantly less favourably in sub-Saharan Africa than in other regions. However, the evaluation synthesis shows that this finding is not robust when we add further control variables (Noltze et al., 2018). Furthermore, supra-regional projects are rated significantly more positively in all areas than regional projects; the only exception to this is the area 'context'. This may be due either to synergy effects between the various programmes, or to the possibility of a 'more holistic' approach compared to individual programmes (or both). For instance, the situation in neighbouring countries might generate effects here, or a wider range of stakeholders might be involved in the process.

[16] *Since the findings on criteria that are mentioned by only very few projects (fewer than 5 per cent of the sample size) possess low explanatory power, we did not make any further use of these criteria in the analysis.*

Figure 6: Percentage of evaluation reports referring to evaluation criteria and areas

Source: Authors' own graphic.

Notes: The graphic shows the percentage of evaluation reports that refer to at least one differentiated criterion for the respective sustainability criteria when assessing sustainability. The project figures shown in blue indicate the percentage of all reports that have reported on at least one criterion for the area. N = 513.

4.2.2 Context

A first and, as expected, important area in the assessment of project sustainability is the context. In accordance with the analysis grid for this meta-evaluation, we included in the analysis those contextual factors that according to reports have a direct effect on results or the predictability of the continuation of results. We disaggregated the contextual factors to take account of the social, economic, environmental and political dimensions. The majority of the evaluations (59 %) include the context when assessing project sustainability (see Figure 6 in Section 4.2.1). Here the reports tend to focus on political aspects. Just under half of all evaluation reports include this criterion (see Figure 11 in the Annex). While economic contextual factors are addressed in around a quarter of all reports, social and environmental contextual factors rarely play a role in the assessment of sustainability. The KfW includes contextual factors in its reports more often than the GIZ (see Figure 13 in the Annex), particularly with regard to economic aspects (see Figure 14 in the Annex)[17]. This might be due to structural differences between TC and FC projects. FC projects usually manage without being present on the ground. At the same time, in some cases considerable amounts of funding are made the responsibility of partners and executing agencies. Hence the context is very important, and due account is taken of this later on in the ex-post evaluations.

The positivity or negativity of contextual effects on the assessment of sustainability is also important. Compared to other factors, contextual factors are largely seen as having a negative effect on project sustainability. This means, for example, that a certain political trend – for instance in the run-up to key elections – is seen as creating uncertainty when assessing sustainability, and ultimately the score is adjusted downwards. A synopsis provides a clear picture: Overall, contextual factors are a critical area in the assessment of sustainability. The high negative difference in the area 'context' as a whole (see Figure 7 in Section 4.2.1) is due particularly to the perceived negative effect of social (and economic) aspects (see Figure 8 in Section 4.2.1): Within this difference, just under 90 per cent (or 70 per cent) of the evaluations that report on this criterion reach a negative assessment. This is constant across all sectors. Only evaluations in the health sector reach a relatively balanced assessment of the effect of contextual factors on sustainability (see Figures 21 and 22 in the Annex).

When evaluating results and their sustainability, the timing of measurement is crucially important. Ex-post evaluations, which make their observations at some point after the end of a project, therefore play an important role. A comparison of ex-post evaluations (by GIZ and KfW) on the one hand, and the other types of evaluation used on the other (GIZ PPRs, PEs and final evaluations) on the other, shows that those evaluation types employed at a relatively early point in time see environmental contextual factors in a significantly more negative light (see Figure 19 in the Annex). By contrast, in the ex-post evaluations there are more cases where the ecological context is assessed as having a positive effect, even though the overall assessment is still largely negative. One possible explanation is that positive results in the environmental dimension only occur after a prolonged period and can therefore only be measured relatively late. There are also differences between the implementing organisations. GIZ's ex-post evaluations see environmental contextual factors in a more critical light than KfW's ex-post evaluations (see Figure 19 in the Annex). This finding points to systematic differences in the assessment of sustainability depending on the type of evaluation. Section 4.1 showed that the evaluation types differ not only with respect to the timing of evaluation but also in terms of quality. The possible effect of quality on the assessment of sustainability therefore requires special attention, and will be discussed at the end in Section 4.3.

Overall, according to the evaluation reports contextual factors have a major effect on the assessment of project sustainability. It also emerged that they are seen largely as having a negative effect. However, this involves a risk of systematic distortion in reporting. It could be that negative contextual effects tend to be mentioned more readily when data are being gathered for evaluations, and as a result occupy a more prominent position in the presentation of the findings, whereas a neutral or positive context is less likely to be mentioned. This is why the accompanying evaluation synthesis integrated into the causal analysis further contextual factors that were not emphasised

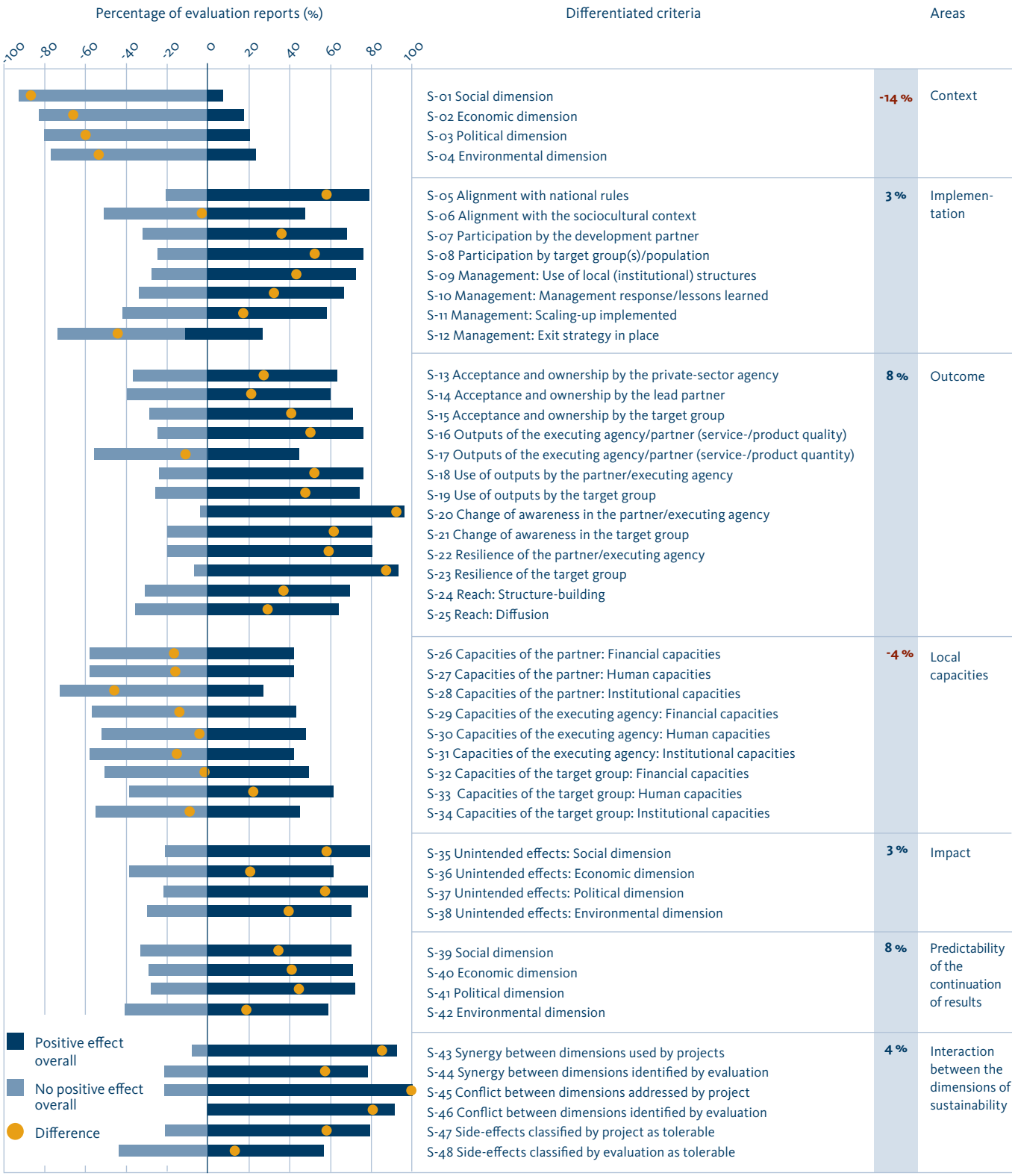
[17] *Since different numbers of GIZ and KfW evaluation reports were included in the analysis (GIZ: n = 341, KfW: n = 172), the frequencies in these two graphics were corrected to take account of these different numbers. Here, 100 per cent of the scale therefore means 100 per cent of the GIZ reports or 100 per cent of the KfW reports.*

Figure 7: Effect of sustainability criteria and areas on the assessment of sustainability

Source: Authors' own graphic.

Notes: The bars show the percentage of evaluation reports that ascribe to the respective sustainability area a positive or negative effect on the sustainability of a project. The entire length of the bar represents in each case 100 per cent of the evaluations reporting on the criterion in question. The bars to the right (and left) of the axis represent the number of evaluation reports that ascribe to the criteria a positive (or negative) effect on sustainability. The dots represent the difference between the percentage of positive and the percentage of negative assessments of a criterion. The percentages shown in blue indicate the average values per area. N = 513.

Figure 8: Percentage of evaluation reports by differentiated sustainability criteria and effect on sustainability assessment



Source: Authors' own graphic.

Notes: The bars show the percentage of evaluation reports that ascribe to each differentiated sustainability criterion either a positive (dark blue) or negative (light blue) effect on the sustainability of a project. Individual differentiated criteria include only those reports that use the differentiated criterion in question to assess sustainability. The dots represent the difference between the percentages of positive and negative assessments of a differentiated criterion. The percentages shown in blue indicate the average values per area. N = 513.

in the evaluation reports. Here it emerged that particularly in ex-post evaluations, there is a significant negative link between the general income situation in a country at the sustainability score awarded for project.[18]

4.2.3 Implementation

The second area of sustainability assessment that we analysed involves aspects of implementation. As per the assessment grid, this includes the criteria 'alignment', 'participation' and 'management'. It emerges that overall, criteria in the area of implementation play a moderate role in the assessment of project sustainability. Although 59 per cent of evaluations include at least one criterion from this area in their assessment (see Figure 6 in Section 4.2.1), in themselves the individual criteria are associated with sustainability only relatively infrequently (see Figure 13 in the Annex).

The criterion' alignment' has been of key importance in the aid effectiveness debate for quite some time. It is usually invoked as a precondition for acceptance and ownership by partners, executing agencies and target groups, and thus as an elementary component of development effectiveness (Hartmuth, 2004; Klingebiel, 2013; OECD, 2017). In this analysis, as in the evaluation reports, this is understood to mean the alignment of a project with local structures – in other words alignment with national development strategies or alignment with the sociocultural context of the target groups. According to the findings of this meta-evaluation, however, the results chain for acceptance and ownership, which extends from the use of outputs, through results, and ultimately on to sustainability (understood as the continuation of development results over time), appears to be relatively long. At least, only about one in ten evaluation reports directly link 'alignment' with the sustainability of projects. When they do, though, they do so largely in a positive sense. Seventy percent of the evaluation reports that report on the criterion alignment see it as a factor for success.

'Participation' has also played a pivotal role in the aid effectiveness debate for quite some time, and is therefore one of the criteria for evaluating development cooperation (OECD, 2010b). As well as the degree of participation, it is also possible to distinguish between the various stakeholder groups. In the analysis grid for this meta-evaluation, participation was included when according to the reports the partners or target groups were at least consulted, and this was important for project sustainability. It then emerged that participation tends to be included relatively infrequently in the assessment of sustainability, though when it is, it is in most cases described as a factor for success. With regard to participation at the level of target groups, the GIZ reaches significantly more positive assessments than the KfW (see Figure 15 and Figure 16 in the Annex). The findings also vary according to type of evaluation. GIZ and KfW ex-post evaluations assess the effect of participation on project sustainability neither positively nor negatively, while GIZ's decentralised and final evaluations reach a clearly positive assessment (see Figure 17 and 18 in the Annex). This indicates that evaluations which take place either during or shortly after the end of the project systematically rate the importance of participation more positively than evaluations conducted some time after projects have been completed.

The final part of our analysis of implementation involved the importance of management-related assessment criteria. Here we analysed to what extent the use of local structures in management, in the management response to monitoring and evaluation recommendations, and in the formulation of scaling-up and exit strategies, was considered relevant when assessing sustainability. Here it emerged that just under half of all evaluation reports included at least one of these four criteria when assessing sustainability (see Figure 6 in Section 4.2.1). Regarding management, these criteria were included in the assessment both as factors for success and as factors for failure, though overall they were usually described as positive. One exception is the assessment criterion 'exit strategy'. The majority of evaluation reports saw this criterion as problematic for sustainability. However, this link is found almost exclusively in GIZ evaluations. KfW evaluations present a relatively balanced view of this criterion (see Figure 17 in the Annex). Seen in the light of the different approaches to implementation, this seems plausible. Given GIZ's strong presence on the ground, its projects are more dependent on phasing-out strategies that work. In FC projects, the handover

[18] *To analyse the context more broadly here we used the current figure for per capita gross domestic product (GDP) in US dollars, the net receipt of ODA transfers as a percentage of GDP and the Freedom House Index. The Freedom House Index provides information on the scope of political rights and civil liberties in a society (Freedom House, 2016).*

arrangements are usually already specified in the module proposal. Based on the data available to this meta-evaluation, however, it is not possible to explain definitively whether the predominantly negative link to sustainability is due to the absence of an exit strategy, or rather to the fact that existing exit strategies were poorly designed. Van Tulder and Pfisterer (2008) also emphasise the major importance of well-executed exit or phasing-out strategies for project sustainability.

To summarise, aspects of implementation are associated with project sustainability in a positive sense in most cases. Differences are evident in the assessment practices of the two implementing organisations, which can be explained by the different implementing structures of TC and FC. Furthermore, in the area of implementation we found no differences of any significance between sectors or regions.

4.2.4 Outcome

The direct and indirect, and short-term and medium-term, results of a development project form a further aspect of sustainability assessment. In the assessment grid we place them together under the heading 'outcome', which we broke down into a large number of different criteria. The most frequently used criteria in this area include 'acceptance and ownership', 'outputs of the executing agency/partner', 'use of outputs' and 'reach' (see Figure 6 in Section 4.2.1 and Figure 11 in the Annex). Just under half the evaluation reports we analysed mention these criteria. Other aspects of this area such as 'change of awareness' and 'resilience and adaptability' tended to be included in the assessment of sustainability infrequently.

In the area 'outcome' we first of all analyse the role of 'acceptance and ownership'. Both these concepts have always been part of the aid effectiveness debate, and are associated with sustainability accordingly (OECD, 2008). The assumption is that acceptance and ownership are prerequisites for successful development cooperation and the continuation of development results over time (Russ-Eft, 2014; Stockmann and Silvestrini, 2011). As the two concepts are mentioned in close connection with each other in the reports, we treated them as a single criterion in this meta-evaluation. We analysed the extent to which the evaluations included the initiative of local actors in their assessment of sustainability. We analysed the concepts separately for the groups 'partners', 'implementing agencies' and 'target groups'. We found that 'acceptance' and 'ownership' are linked to sustainability in approximately one in every two evaluations (see Figure 6 in Section 4.2.1), and that they are included in the assessment of sustainability as success factors in most cases. Across all types of evaluation, GIZ sees acceptance and ownership in its projects in a significantly more positive light than KfW.

A further aspect of our analysis was the direct outputs of projects. We analysed the extent to which the quality and quantity of outputs were assessed as being sufficient to achieve the project objectives. We found that the quality of outputs was included in the sustainability assessment significantly more frequently than their quantity (see Figure 11 in the Annex). Furthermore, the quantity of outputs displays a negative difference – albeit a slight one (see Figure 8 in Section 4.2.1). This may mean that although quantity is less important than quality when assessing sustainability in evaluation reports, the absence of a certain quantity may nevertheless have a negative effect on project sustainability.

Building on these findings, the meta-evaluation analysed the extent to which the generation of outputs also entails their use. When analysing the 'use of outputs' we distinguish between use by partners and/or executing agencies, and use by the target groups. We found that just under one in four reports include the use of outputs in the assessment of sustainability (see Figure 6 in Section 4.2.1). While GIZ evaluations report chiefly on the use of outputs by partners/executing agencies, KfW evaluations more often mention the use of outputs by the target group (see Figure 14 in the Annex). Here too, one possible explanation would be the different implementing structures. Many TC outputs are generated by local personnel and are designed initially for the partners or local implementing structures, which in turn generate outputs for the target groups. By contrast, FC outputs are generated by the local executing agencies and delivered directly to the target groups. The use of direct inputs is then in most cases included as a factor for success when sustainability is assessed. Here too, differences between the implementing organisations are evident. GIZ evaluations see

the link between the use of output and sustainability in more positive terms than KfW evaluations (see Figure 15 and Figure 16 in the Annex).

In the area 'outcome' we also looked at whether and to what extent projects contributed to sustainability at the level of partners/executing agencies/target groups through a 'change of awareness'. In this connection we analysed the extent to which the evaluations referred to long-term behavioural changes among the actors concerned. Our assumption was that changes in the actors' awareness would have a particularly strong effect on project sustainability (Stadtler, 2016; Von Raggamby and Rubik, 2012). At first glance the findings confirm this assumption only to a limited extent. Only 15 per cent of all the evaluations referred to changes in awareness and behaviour among target groups when considering project sustainability (see Figure 6 in Section 4.2.1). However, when they do report doing so their assessment is a highly positive one. With 70 per cent of all evaluations reporting having considered 'change in awareness', this criterion displays the highest aggregate difference (positive-negative difference) of all the outcome criteria. This therefore confirms the assumption that changes in awareness have a relatively strong effect on the assessment of sustainability. With regard to the executing agency/partner, across all evaluation reports the GIZ sees change in awareness as having a highly positive effect (see Figure 15 in the Annex). The KfW evaluations also see positive effects, although the aggregate difference is much more balanced – the figure is around 30 per cent of the KfW evaluations reporting on changing awareness.

This meta-evaluation also analysed the importance of 'resilience and adaptability'. Referring to the evaluation reports, we analysed the extent to which projects enabled the partners/executing agencies and/or target groups to self-reliantly identify development potential and risks, and translate this into action. We found that the resilience and adaptability of actors was only rarely included in the assessment of sustainability. Only around one in five evaluation reports mention this criterion (see Figure 6 in Section 4.2.1). Having said that, resilience is largely seen as a success factor for sustainability. Particularly the resilience of target groups is described as conducive to success (see Figure 7 and Figure 8 in Section 4.2.1). One surprising finding is that only in the education sector is resilience and adaptability seen as having a largely constraining effect on sustainability (see Figure 21 in the Annex). Here we had assumed that education projects would be conducive to target group resilience. In the present study we were unable to substantiate this empirically.

Finally we analysed the importance of 'reach' in the assessment of project sustainability. Here we focused on what the evaluation reports had to say about the criteria 'structure-building' and 'diffusion'. Using the criterion of structure-building we looked at the extent to which changes had occurred at the system level that were also used to assess sustainability. With regard to diffusion, we looked at the extent to which outputs and innovations had been disseminated beyond the original target group. Based on the literature, we assumed that reach would be a significant success factor for project sustainability (Stadtler, 2016; Vahlhaus, 2014; Von Raggamby and Rubik, 2012). Our findings fully corroborate this assumption. Reach is linked to sustainability in over half of all the evaluations, and is thus one of the most frequently reported criteria in the area 'outcome' (see Figure 6 in Section 4.2.1). On closer inspection it emerges that the criterion 'structure-building' is significantly more important than the criterion 'diffusion'. GIZ evaluations in particular mention structure-building frequently when discussing sustainability (see Figure 11 and Figure 14 in the Annex). Once again, a possible explanation for this is structural differences between TC and FC projects. While FC projects use resources largely to build infrastructure, and capacity building with target groups and disseminators usually takes place only as an 'accompanying measure', capacity building is a core component of TC projects.

A synopsis shows that according to the evaluation reports, a number of assessment criteria in the area 'outcome' appear to have a clearly positive effect on project sustainability. This is the case inter alia with the criteria 'change of awareness' and 'resilience'. Particularly the GIZ evaluations establish clear positive links here. A specific sectoral feature is evident in evaluations of projects in the transport sector. Here, the effect of outcome criteria of sustainability is seen in a much more

negative light. This finding might be linked to the fact that the success of outputs generated in the transport sector is heavily dependent on demand. One demand-based indicator here would be economic activity in the project setting.

4.2.5 Local capacities

Another area in which project sustainability is assessed is local capacities. The term 'capacities' is used here to mean the financial, human and institutional contributions made by the partners, executing agencies and ultimately also the target groups. 'Local capacities' is used as a generic term for the ability of local actors to continue the outputs and maintain the results over time. In the literature, major importance for sustainability is ascribed to local capacities (Caspari, 2004; KfW Entwicklungsbank, 2003; Russ-Eft, 2014; Stockmann and Silvestrini, 2011). Based on our empirical results we can confirm this importance. Eighty-six per cent of all evaluations include local capacities in their assessment (see Figure 6 in Section 4.2.1). The capacities of executing agencies are linked to sustainability the most frequently. This high value results chiefly from the ex-post evaluations of the KfW. KfW project usually work through local executing agency structures, and therefore ascribe major importance to them in evaluations.

According to the evaluation reports, local capacities in most cases are seen as having a negative effect when assessing sustainability. Presumably this is to be explained by the insufficient capacities of partners, executing agencies and target groups in the partner countries of German development cooperation. Nonetheless, this finding is astonishing in that insufficient capacities must be taken into account when projects are planned, and should therefore be the focus of attention long before any evaluation takes place. What is particularly surprising is the fact that the capacities of partners are seen as a significant challenge when assessing sustainability. These, however, are negotiated as part of project agreements, and should therefore be much easier to plan than for instance the contributions of target groups, which are reported much more frequently as being a factor for success (see Figure 7). From the perspective of projects there is a need to clarify the extent to which the assessment of partner capacities can be approved during project planning, in order to prevent negative effects on sustainability later on. Ultimately, a sound analysis of local capacities is also in the interests of evaluations, as these can quickly find themselves being accused by partners of shifting responsibility for the success of projects onto external factors.

It is also interesting to compare the GIZ and KfW here. In their evaluation reports the two organisations reach significantly different conclusions regarding executing agency capacities. While the GIZ sees executing agency capacities as problematic for project sustainability, the KfW takes a more positive view overall. One possible reason for this is the more sophisticated ex-ante appraisals for FC projects, which ultimately identify reliable partners. By contrast, in TC projects the thematic focus, such as good governance, not infrequently involves working with executing agencies that have a high capacity development support requirement. This assumption is corroborated by our sectoral analysis, where we see that the sector 'peace' receives, all things considered, by far the most negative sustainability assessment for the criterion 'executing agency capacities'.

In other words the differentiated analysis confirms the overall impression that local capacities display only a slight difference in assessment, even though moderate differences are to be observed between the various groups and criteria. Overall the GIZ reaches more negative assessments than the KfW, and the ex-post evaluations deliver more positive assessments than the GIZ's decentralised evaluation types. As in the two preceding areas of sustainability – 'implementation' and 'outcome' – here too sub-Saharan Africa turns out to be the region where local capacities are assessed as having a significantly more negative effect on sustainability than is the case in other regions (see Figure 24 in the Annex). Projects in Latin America and Europe/Caucasus also see local capacities in a negative light, though to a much lesser extent.

4.2.6 Impact

A further key area in the assessment of sustainability is impact. Here we assumed that projects which contribute to impact would be more successful and sustainable than projects that generate only direct outputs (Boone, 1996; Faust, 2007). In order to make transparent appropriately the importance of impact for project sustainability as reflected in

the evaluation reports, this meta-evaluation analysed the findings on impact comprehensively. This involved 1) comparing intended with substantiated results at the level of overarching objectives by dimension, and 2) including the findings on unintended effects in relation to the dimensions.

Our analysis of the sample shows that social and economic overarching objectives were specified in around 60 and 50 per cent of projects respectively, making these the most frequent categories. Political and environmental overarching objectives were pursued less frequently. Many projects have overarching objectives in more than one dimension of sustainability.[19] When comparing the overarching objectives we found that the evaluation reports rate the achievement of project objectives very positively (see Figure 9). This remarkable degree of achievement of objectives is evident across both implementing organisations and all evaluation types. Only the sectoral analysis reveals minor differences. The sector 'peace', for instance, displays a relatively low degree of the achievement of objectives. This is consistent with the observation made in this study that findings in this sector depend heavily on contextual factors. By contrast, other sectors display a relatively high rate of achieving objectives, with examples including 'democracy', 'economy' and 'energy' (see Figure 27 in the Annex).

While the intended results are included in the majority of evaluations in relation to the overarching objectives, unintended effects are barely discussed at all. Only about one in five reports mentions a positive or negative unintended effect (see Figure 6 in Section 4.2.1). Comparison of the positive and negative significance of unintended effects for sustainability assessment produces a significantly positive picture. Seventy per cent of evaluations that mention unintended effects link this to project sustainability in a positive way (see Figure 7 in Section 4.2.1).

One striking feature is the assessment of economic aspects compared to social, political and environmental aspects. Seventy to eighty per cent of the evaluations consider the unintended social effects to be positive. In other words the difference is positive and the figure is 50 to 60 per cent of the evaluations. This is driven inter alia by the GIZ, which sees the criterion of 'social side effects' in a significantly more positive light than the KfW (see Figure 15 in the Annex). Economic aspects, on the other hand, are seen significantly less favourably. Forty per cent of the evaluations reporting on this, i.e. a not insignificant percentage, conclude that unintended effects have a negative effect on sustainability (see Figure 6 in Section 4.2.1).

4.2.7 Predictability of the continuation of results

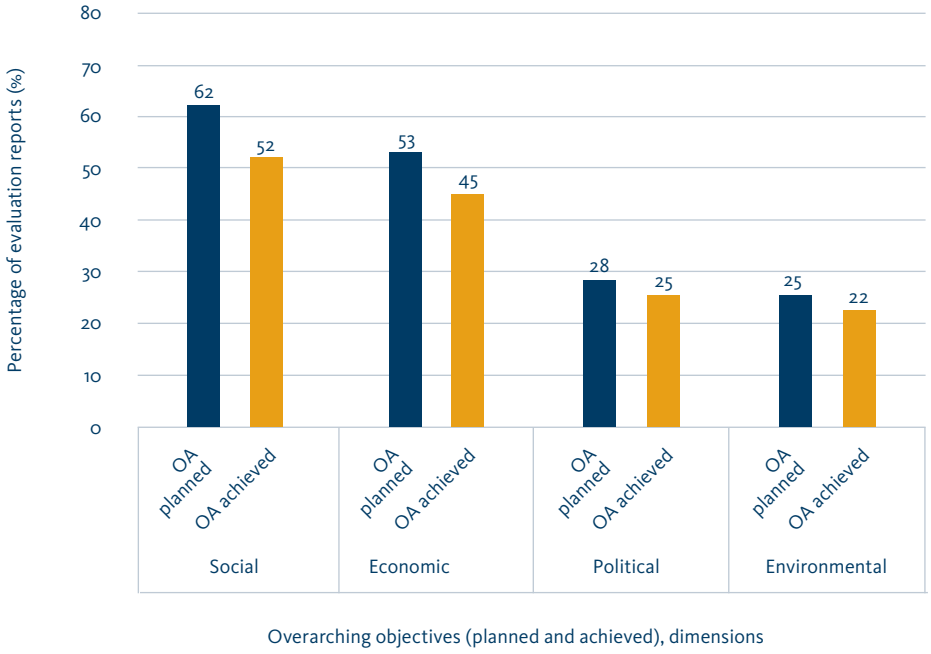
The predictability of the continuation of development results over time is a key aspect of assessing project sustainability. On a purely conceptual level, predictability is even the key aspect when assessing 'durability'. We included this aspect in the meta-evaluation's analysis grid with respect to the achievement of overarching objectives over time. It is therefore surprising that only one out of two evaluation reports explicitly discusses the predictability of the continuation of results over time. One obvious explanation here is the fact that shortcomings in the substantiation of results mean that not all evaluations are able to draw definitive conclusions concerning the achievement of the planned overarching objectives. Logically, in these cases it is then not possible to make any assessment of the predictability of the continuation of results. In accordance with the analysis grid, we also analysed the predictability of the continuation of results in relation to the various dimensions of sustainability. Since the overarching objectives of GIZ and KfW projects can in most cases be assigned to the social and economic dimensions (see Section 4.2.6), we will discuss the predictability of the continuation of results over time chiefly within these two dimensions (see Figure 11 in the Annex).

As a rule, the evaluations see the predictability of the continuation of results as a positive factor when assessing sustainability (see Figures 7 and 8 in Section 4.2.1). This finding remains constant across both implementing organisations and the various evaluation types (see Figures 15 to 18 in the Annex). One exception is environmental results, which according to the findings of the ex-post evaluations jeopardise the sustainability of projects. One possible explanation for this is that results in the environmental dimension are first of all

[19] *This is discussed in Section 4.2.8 on the 'Interaction between the dimensions of sustainability'.*

more difficult to achieve, and secondly difficult to maintain. We also identified a specific sectoral aspect. In the energy sector, the predictability of the continuation of results is seen in a significantly more negative light. This is also the only sector in which this factor is seen as having a negative effect on sustainability (see Figures 21 and 22 in the Annex). This could be due to the fact that in this sector an important role is played particularly by infrastructure projects, in which it is particularly difficult to assess whether necessary maintenance work is certain to take place over time.

Figure 9: Percentage of evaluation reports by planned and achieved overarching objectives, and dimension of sustainability



Source: Authors' own graphic.

Notes: The graphic shows the percentage of evaluation reports by overarching objectives that were planned (blue) and achieved (orange). N = 513.

4.2.8 Interaction between the dimensions of sustainability

Pursuant to the BMZ guideline of 2006, evaluations should analyse and describe both results and the predictability of the continuation of results in relation to the social, economic, environmental and political dimensions. Since the guideline does not explicitly require analysis of the interaction between these dimensions, this meta-evaluation enquires whether a discussion on possible interactions nevertheless did already take place in the past. Our motivation in systematically pursuing this issue results from the prominent position that the 2030 Agenda accords to interaction between the dimensions of sustainability.

When analysing 'potential effects and challenges between the dimensions', we looked at whether the evaluation reports had emphasised synergies between the dimensions, mentioned any conflicting objectives, or concluded that possible side-effects in the individual dimensions were tolerable. We also examined whether the evaluations confirmed or refuted these assessments. Compared to the other areas, interaction between the dimensions so far been made part of the assessment of sustainability significantly less frequently. Only about a quarter of the evaluations mentioned such aspects. Synergies between the dimensions were mentioned the most frequently (see Figure 6 in Section 4.2.1 and Figure 11 in the Annex). Furthermore, the assessment practices of the two implementing organisations differ. GIZ evaluations mention synergies more often, whereas the KfW consultants refer more frequently to conflicts between the dimensions. At first glance this finding seems plausible, as for instance the construction of infrastructure, as is often the case in FC projects, can damage the environment or impact negatively on neighbouring communities. For TC projects a link of this kind would appear less obvious in the first instance, as their activities usually revolve around capacity building. Not least for this reason, the KfW prescribes a corresponding impact assessment in its sustainability guideline (KfW Entwicklungsbank, 2016). That said, conflicts between the dimensions can also arise in TC projects. For example,

promoting ownership without the right political support – such as regulations to limit activities, compliance with which is then guaranteed through inspections – can lead to increased environmental burdens. More comprehensive reporting can therefore help bring about more coherent project planning and implementation.

Where synergies between the dimensions were promoted by projects, and this was subsequently confirmed by the evaluation, in almost all cases these were seen as positive factors when assessing sustainability (see Figures 7 and 8 in Section 4.2.1). Overall, however, evaluators hardly ever examine this criterion when assessing sustainability. This failure to include interaction between the dimensions, in conjunction with the inadequate analysis of unintended effects, constitutes the second key deficit in current evaluation practices with regard to the requirements arising from the 2030 Agenda.

4.3 Links between evaluation quality and the assessment of sustainability

This meta-evaluation has demonstrated that German development cooperation evaluates the sustainability of its projects comprehensively. With respect to the conceptual framework of the DAC criteria, sustainability is thus a comprehensive and overarching construct that extends far beyond the 'evaluation criterion sustainability'. However, the findings of the preceding sections lead us to assume that both the number of criteria applied, and the assessment of whether individual criteria have a negative or positive effect on project sustainability, are also dependent on the particular type of evaluation, and are therefore determined not only by the underlying understanding of sustainability. In the chapter on findings, we demonstrated that the types of evaluation display differing levels of evaluation quality. We will now address the issue of whether and to what extent the methodological quality of evaluations actually affects the assessment of sustainability.

With respect to the basis on which project sustainability is assessed, two different links are important: 1) the link between evaluation quality and the number of criteria that the report addresses, and 2) the link between the quality of evaluation and the tendency to see particular criteria as having more of a positive or negative effect on sustainability.

Here we see a positive link between quality and breadth of criteria as a basis for assessment (see graphic on left in Figure 10). This means that evaluations of high quality tend to include more sustainability criteria in their assessment. In other words, the assessment of sustainability is based on a broader foundation. With regard to the appropriate application of a comprehensive understanding of sustainability, including a wide range of criteria might well be conducive to a robust overall assessment of sustainability. However, the question also arises of whether quality entails changes not only in the robustness, but also in the assessment of the individual criteria and ultimately also in the sustainability score. These anticipated changes were not confirmed (see graphic on right). The meta-evaluation finds no statistical link between the quality of an evaluation, and the positivity or negativity of the effects of the assessment criteria. In other words, the methodological quality of an evaluation has no positive or negative effect on the assessment of sustainability. Furthermore, the accompanying evaluation synthesis also found no link between the quality of an evaluation and the sustainability score (Noltze et al., 2018). A wide range of assessment criteria thus only broadens the empirical basis of the analysis and the potential for learning, but has no effect on the final score.

4.4 The evaluation of sustainability by international comparison

Part of this meta-evaluation was devoted to the contextual study of the appropriateness of German evaluation practices, which included a comparative international perspective. This involved looking at how 40 evaluation units of the OECD-DAC EvalNet and nine multilateral organisations deal with sustainability as an evaluation criterion (see Section 3.4). As well as information published on the official websites of the units, we also included in the study standards and guidelines etc. that were available online.

Figure 10: Quality index by number of differentiated sustainability criteria and by aggregate effect on the assessment of sustainability

Source: Authors' own graphic.

Notes: The graphic on the left shows the link between the quality index of a report and the number of differentiated criteria used to assess sustainability. The graphic on the right shows the link between the quality index of a report and the aggregate effect (or 'positive-negative difference') of all differentiated criteria used to assess sustainability. The aggregate effect is calculated as the sum of all the differentiated criteria assessed in the report as positive (+1), neutral (0) or negative (-1). N = 513.

The contextual study identified a low overall transparency of evaluation practices for sustainability. Only 18 of the 40 DAC EvalNet units and six of the nine multilateral organisations publish transparent information on their websites concerning how they deal with sustainability in project and programme evaluations. For the remaining organisations one can only assume that sustainability is applied to measure the success of projects as one of the five DAC criteria.

The information available on the remaining 24 evaluation units concerns largely the understanding of sustainability applied in project and programme evaluations. These definitions of sustainability revolve around the continuation of development results over time. However, the understanding of sustainability is usually broken down by dimension of sustainability. Across all 24 evaluation units, based on the frequency of mention we identified the following ranking: 1) financial, 2) institutional, 3) political, 4) social, 5) technical and 6) environmental sustainability.

As well as the conceptual definition of sustainability, a further key focus of interest in the contextual study was the operationalisation of the criterion of sustainability for project evaluation purposes. We studied whether and to what extent the evaluation units concretised their conceptual understanding of sustainability by employing verifiable assessment criteria. Our comparative analysis revealed that only few evaluation units explain transparently how they operationalise sustainability as an evaluation criterion. When they do, one of the common criteria is a risk assessment in the

project context. Beyond that, sporadic mention is made of principles from the aid effectiveness debate, such as ownership, being used as additional assessment criteria. By publishing the BMZ guideline on applying the DAC criteria in 2006, German development cooperation did succeed in providing a condensed description of the assessment criteria in a single paper. By international comparison this remains peerless.

Finally, we looked at whether the evaluation units studied use not only specific assessment criteria, but also systematic rating scales such as points or scoring systems. It emerged that apart from Germany, only few countries also prescribe rating systems. These include Japan, Switzerland and France. Among the multilateral organisations, such standards are more common. As well as the United Nations Development Programme (UNDP), the multilateral banks – the European Investment Bank, the World Bank and the African and Asian development banks – also use both criteria and scoring systems.

The findings demonstrate that despite the basically standardised use of the strict understanding of sustainability as the continuation of development results over time, by comparison across the OECD sustainability is often applied as an evaluation criterion significantly more broadly, though conceptually inconsistently. One possible reason for this is the cross-cutting nature of the criterion sustainability for all DAC criteria – conceptually, the prerequisites for sustainability are linked to the other DAC criteria. This low level of conceptual harmonisation of sustainability goes hand-in-hand with a low level of standardisation of assessment practice. Only few of the evaluation units studied operationalise their understanding of sustainability transparently by prescribing specific assessment criteria and rating systems. This low level of standardisation also means that there are few quantitative cross-section evaluations of sustainability in development cooperation. Alongside a small number of national evaluation units, it is chiefly the international development banks that create the conditions for aggregating knowledge through overarching analyses. Germany too provides an enabling environment for quantitative cross-section evaluations. In 2014, however, the GIZ switched from its four-point rating system to a point scale, which as part of overall performance measurement is being converted into a six-point rating scale, whereas the KfW continues to award scores along a four-point scale.

5.

CONCLUSIONS AND RECOMMENDATIONS

This meta-evaluation forms the first systematic and comprehensive empirical survey of practices for evaluating sustainability in German development cooperation projects. Sustainability has long been a guiding principle in German development cooperation. Its growing importance in international cooperation makes this meta-evaluation especially relevant. The 2030 Agenda has made the principle of sustainability now more than ever before *the* guiding framework for the strategic and operational orientation of development cooperation.

The key purpose of this meta-evaluation is to support development cooperation in developing modern evaluation practices that are geared to sustainability. To do so it ventures into uncharted methodological territory. Beginning with a traditional design, the meta-evaluation goes beyond merely assessing the methodological quality of evaluations to systematically analyse the assessment criteria they apply. This was the only way to systematically analyse both the underlying conceptual understanding of sustainability, and the evaluation practices based on that logic. Using an integrated research design framework, the findings of the meta-evaluation were subsequently used as part of the accompanying evaluation synthesis to study the factors affecting sustainability (Noltze et al., 2018).

5.1 The quality of German evaluation practice

The first part of the meta-evaluation is devoted to the quality of evaluation practices. As well as the quality of individual evaluations, a further object of study was the structures of the underlying evaluation system. The conclusions below therefore concern both the quality of individual evaluations and types of evaluations, and the wider framework of evaluation in German development cooperation. Accordingly, the recommendations therefore relate first of all to the further development of evaluation practices, and then to the further development of the evaluation system.

The findings demonstrate that the quality of the findings and conclusions obtained by the GIZ and KfW from their module evaluations are appropriate for evaluations of that size. As well as describing the object of the evaluation, a large majority of the reports include a logical description of the causal links to be analysed and the methodological approach. German development cooperation is characterised by a high degree of coverage. The GIZ subjects almost all its modules to systematic evaluation. The KfW works with a representative sample; every year, half of all the evaluation-ready projects per sector are subjected to an ex-post evaluation.

The meta-evaluation demonstrated that there is potential for improvement regarding evaluation quality. First of all, great effort should be made to detect causal relationships by applying systematic methods of analysis and triangulation. Secondly, the logic of findings and conclusions should be made more transparent. A more robust and logically transparent substantiation of results is key to reliably demonstrating sustainability in accordance with the principles of the 2030 Agenda. As well as the purely methodological options, further potential also exists regarding the appropriate timing of data collection. Particularly in the GIZ's many decentralised evaluations, which are conducted in the course of projects, the substantiation of results is based entirely on assessments of the future, which inevitably involves uncertainty. By contrast, ex-post evaluations offer an opportunity to actually observe results and the sustainability of results after a certain interval following completion of the project.

The module evaluations conducted by the KfW and GIZ usually involve comparing actual values with target values for selected indicators drawn from the results logic. Although such comparisons do not completely close the attribution gap, they do allow an approximate identification of causal relationships. Given that this is the case, it is astonishing that only few evaluations indicate that they make use of monitoring data supplied by projects or executing agencies. This runs counter to the logic of robust comparison of actual values with targets.

The evaluation team understands that the quality of the decentralised evaluations is difficult to improve, given the fact that the evaluation missions are overstretched. In the majority of the GIZ's decentralised evaluations, the purpose is to focus not only on the DAC criteria, but also on management-related aspects – a fact that is also motivated by the need to prepare a

proposal for a possible follow-on phase. Consequently, decentralised evaluations are more like appraisals than evaluations. Apart from possible inherent conflicts of interest that can arise between appraisal and evaluation reports, the needed resources also need to be taken into account – regardless of whether the two purposes are pursued separately or jointly.

Recommendations on further developing evaluation practice

1. 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 existing potential to increase the quality of evaluation, particularly with respect to substantiating results and sustainability.
2. Bearing in mind the low importance persistently ascribed to monitoring data in module evaluations, the implementing organisations should systematically examine what obstacles exist here and how these can be overcome. In this context they should examine whether project monitoring systems can be linked through their objectives systems to the system of goals and targets that make up the Sustainable Development Goals (SDGs).
3. To ensure transparency and incentivise clear reporting the GIZ and KfW should, while remaining mindful of the opportunities and risks, explore the possibility of publishing their evaluation reports in full – perhaps initially in a pilot phase – and informing the BMZ of the lessons they learn in the process.
4. To raise the quality of evaluation, the team recommends that GIZ institutionalise quality assurance in the Evaluation Unit on a long-term basis. In the future, all module evaluations should be managed by the Unit.
5. To help raise the quality of evaluations, the GIZ should separate appraisal and evaluation.
6. Regarding the appropriate point in time at which to reliably substantiate results and sustainability, greater importance should once again be attached to ex-post evaluations. When ex-post evaluations are being conducted, both the GIZ and KfW should ensure that the importance of management is understood. This can involve for instance defining key focuses, or selecting an appropriate point in time for the evaluation.

Recommendations on further developing the evaluation system

7. To promote joint learning and accountability, the team recommends that the BMZ harmonise the practice of evaluation by the GIZ and KfW on the basis of the joint procedural reform (GVR) and the Guidelines for bilateral Financial and Technical Cooperation. In this context the BMZ should issue firm instructions concerning the timing, scope and rating system in order to standardise the types of evaluation for module evaluations.
8. By defining uniform minimum standards the BMZ should support the exhaustive use of potential to raise evaluation quality in module evaluations. The requirements for an evaluation might for instance be concretised by developing a specimen Terms of Reference. Standards might also be introduced at an early point in the process, for instance concerning the requirements included in invitations to tender for evaluation missions (e.g., regular participation in training for evaluators).
9. The BMZ should require the implementing organisations to make their evaluation reports clear and easy to understand, so that they can be read on a stand-alone basis. Depending on the outcome of a corresponding review, the BMZ should require the implementing organisations to publish their evaluation reports in full.
10. The BMZ should ensure that, in addition to the quality assurance of the module evaluations performed by the evaluation units of the GIZ and KfW, an external, cross-organisational meta-evaluation of a random sample of evaluations should be performed on a regular basis.

5.2
Assessing sustainability in German development cooperation

For the first time, the findings of this meta-evaluation demonstrate empirically that in the evaluation of German development cooperation, sustainability is already being understood in a comprehensive sense, and evaluated and assessed accordingly. Bearing in mind the broad debate on the concept of sustainability in development cooperation, this finding perhaps comes as no great surprise. However, it is indeed remarkable in light of the considerably more specific instructions contained in the BMZ guideline on applying the DAC criteria in evaluations. The findings show that the understanding of sustainability applied by evaluation practitioners already goes significantly further than the continuation of development results over time. At the same time, however, it is evident that key elements of the 2030 Agenda, such as the debate surrounding interactions between the dimensions of sustainability, are not yet an integral part of assessment practice, and hence that sustainable development is not yet being fully covered. The findings thus refute the widespread assumption among development professionals that the DAC criteria imply an exclusively narrow understanding of sustainability that is restricted to the continuation of results over time. Yet they also point to a significant discrepancy in relation to the modern understanding of sustainability inherent in the 2030 Agenda.

Moreover, the findings show that in practice sustainability is currently being assessed unsystematically and inconsistently. This is due to the lack of a conceptual framework for a comprehensive understanding of sustainability. So far, selection of the specific assessment criteria has been left very largely to the discretion of the consultants involved. Even the key questions proposed in the BMZ guideline in 2006 are not being applied systematically. Overall, it is evident that the DAC criteria as they stand do permit the evaluation of sustainability understood in a comprehensive sense, but by no means prescribe this on a systematic and binding basis. This lack of a systematic approach means that a simple comparison of sustainability scores across different projects is only possible to a limited extent. This is not conducive to learning from evaluations. At present, a rigorous comparison of the sustainability of projects is only possible at considerable expense and with considerable effort – such as the effort made in preparing the present expanded meta-evaluation and the accompanying evaluation synthesis.

The meta-evaluation has shown that in practice, sustainability is being assessed in relation to a large number of different criteria. As well as the context of projects and local capacities, findings on project outcome are also being used to assess project sustainability. However, criteria are also found which – contrary to the prior assumptions of the meta-evaluation – are used relatively infrequently to assess sustainability. These include criteria on unintended effects and interaction between the sustainability dimensions. The latter is surprising because, although the interaction of results in the different dimensions of sustainability is not yet an explicit part of the guidance provided, it has been part of the development debate for quite some time. By contrast, assessing unintended effects is already a designated part of assessing impact. With regard to the 2030 Agenda, which assigns a prominent role to interaction between the dimensions, there is potential here for further developing evaluation and assessment practices. One reason for the fact that both unintended effects and interaction between the dimensions are barely addressed is the absence of a methodological framework. Here it would be necessary when planning modules to make appropriate provision for evaluation later on. However, neither unintended effects nor interaction between the dimensions are currently being incorporated systematically when logical frameworks for results are formulated. This means that in subsequent evaluations, these effects and results are addressed systematically either at great expense, or virtually not at all. Ultimately, however, the question also arises of what would be the appropriate level of analysis. As development cooperation programmes, and the TC and FC modules that form a part of them, become increasingly complex, many interactions and unintended effects – particularly at the impact level – can only be identified definitively at the level of programmes. In this respect, at the level of individual modules evaluations can only deliver an incomplete picture. Ultimately, the debate on the interaction between dimensions is symptomatic of the need for a debate concerning the large number of evaluation

challenges surrounding the principles of the 2030 Agenda.

The findings of the meta-evaluation also revealed an interesting link between the quality of evaluations and the quantity of information produced. As the methodological quality of evaluations rises, so too does the number of criteria used to assess sustainability. More sophisticated evaluations thus place the assessment of sustainability on a broader footing, and are conducive to the generation of reliable findings. There is, however, no link between quality and the assessment of an individual criterion or the overall assessment of the sustainability of a project.

The analysis of assessment criteria also showed that criteria in the area of project outcome were seen largely as enabling factors for sustainability, whereas criteria in the areas of local capacities and project context were seen largely as constraining factors. On the one hand this finding underlines the challenging framework in which German development cooperation operates. On the other hand, it also entails a risk of externalising responsibility. According to the evaluation reports, low sustainability is caused largely by factors outside the sphere of influence of projects. However, knowledge of difficult frameworks should be available where possible a priori, which would preclude these frameworks having a one-sided effect on the assessment of sustainability later on. In this regard the question also arises of whether potential external risks for the sustainability of German development cooperation projects can be further minimised by improved ex-ante appraisal and planning.

Ultimately these conclusions demonstrate the value of this expanded meta-evaluation, which supplements the assessment of evaluation quality with a discussion of possible risks for sustainability. The design enabled the evaluation team to highlight structural enabling and constraining factors for the assessment of sustainability. The overarching analysis of the object of the evaluation also enabled the team to aggregate knowledge at the global level. Furthermore, the thematic meta-evaluation provided the evaluation synthesis with a broader database (Noltze et al., 2018).

In the future, working with the 2030 Agenda and the sustainability of development cooperation projects in evaluations will be a global task. With respect to German development cooperation, this meta-evaluation has identified a specific need for action. The conclusions presented call for a reform of existing evaluation practices. Alongside the idea of harmonisation and coordination contained in the Paris Declaration and the Accra Agenda for Action, the universal nature of the 2030 Agenda also calls for sharing and coordination at the international level (OECD, 2008; UN, 2015). The recommendations below are thus designed to support the ongoing reform process at the level of German development cooperation, particularly in the context of the joint procedural reform (GVR)[20]. They should also enrich the debates at the international level, particularly within the OECD-DAC. Against the background of the ongoing reform processes, the evaluation team has supplemented the recommendations with a number of conceptual proposals designed to prompt further reflection – in the knowledge that these ideas will be fed into a system of which DEval is a part.

In line with the breakdown of recommendations on the quality of evaluation (Section 5.1), the evaluation team's recommendations on the assessment of sustainability are presented below in two parts: recommendations on further developing evaluation practice, and recommendations on further developing the evaluation system.

Recommendations on further developing evaluation practice:

The recommendations below are addressed to both the BMZ and the implementing organisations. The recommendations should be implemented on the basis of a joint process led by the BMZ and involving the implementing organisations and DEval. The team recommends that this process, including a pilot phase, should be completed by the end of 2018, in order to guarantee from 2019 onwards that evaluation in German development cooperation is in conformity with the 2030 Agenda.

[20] *The BMZ's joint procedural reform (GVR) forms the basis for the future design, implementation and evaluation of country strategies, development cooperation programmes and modules, with the aim of making German cooperation more effective. At various points the GVR refers to the principles and SDGs of the 2030 Agenda. Based on the GVR the implementing organisations are working on organisation-specific guidelines for developing projects, and appraisal and evaluation systems to support them, which are in conformity with the 2030 Agenda. At the GIZ, the in-house evaluation system is being reformed within the framework of an evaluation policy put forward in 2017.*

11. The evaluation team recommends that in the future the BMZ and the implementing organisations should evaluate the sustainability of projects based on the principles of the 2030 Agenda for Sustainable Development, within the framework of an additional assessment criterion.
    - An additional assessment criterion of this kind could be conceptualised such that it supplements the five OECD-DAC evaluation criteria by assessing the contribution to sustainable development as understood by the 2030 Agenda. The additional criterion could be operationalised through appropriate key questions aligned with the structure of the 2030 Agenda principles[21]. The additional criterion would deliver added value for learning and accountability by describing in a condensed manner the specific contributions made by German development projects towards implementing the 2030 Agenda. Furthermore, an additional criterion of this kind would provide the foundation for future reporting on aggregate results in relation to the 2030 Agenda. At the same time the system of DAC criteria could be retained, and with that the comparability with earlier assessments and international harmonisation.
    - Alternatively, key questions based on the principles of the 2030 Agenda could be integrated into the DAC criteria. The advantage of this would be that the DAC criteria would remain the sole basis for assessment. However, their content would be modified and they would then no longer be comparable either historically or internationally. It is also to be expected that integrating a current development agenda would to some extent bring to an end the timeless nature of the DAC evaluation criteria.
    - Regardless of whether the response to the 2030 Agenda in evaluations were to involve a separate criterion or be integrated into the DAC criteria, it would be advisable to discuss this at the international level within the OECD-DAC.
12. As well as including sustainability as conceptualised in the 2030 Agenda as an additional criterion, the BMZ should sharpen the conceptual focus of the DAC criteria and make the BMZ guidelines for applying the DAC criteria more binding.
    - Here it might be possible, allowing for an appropriate degree of case-specific openness while retaining the binding nature of the guideline at an overarching level, to review the key questions with respect to both their genuine nature and the clarity of their conceptual distinction from the key questions for the other DAC criteria, and make them more specific where appropriate.
    - The guideline might then be visualised clearly using an evaluation matrix. Where possible, this might also include proposals for weighting the individual key questions and definitions for intersubjectively comparable scores.
13. As part of the reform of evaluation criteria for assessing the performance of development cooperation projects, the evaluation team recommends that the BMZ retain the existing OECD-DAC criterion of sustainability – understood as implying the continuation of results – and align its key questions with this element.
    - Within the German development cooperation system, some thought should be given to the terminology that would best articulate the conceptual distinction between the continuation of development results, and sustainable development as understood in the 2030 Agenda. Whatever options are selected in the German language, care should be taken to ensure that these are compatible with the international conceptual framework.
14. With respect to the principles of the 2030 Agenda, the GIZ and KfW should investigate how in future evaluations they can identify and assess the unintended effects of a project and the interactions between the dimensions of sustainability.
    - Here it might be possible to describe and analyse anticipated and actual synergies and conflicts between development objectives. Responsibility would begin at the project planning stage. Module proposals could already discuss unintended effects and interactions as elementary components of integrated approaches. Mainstreaming of this kind would require a corresponding directive from the BMZ.

[21] *The principles include: shared responsibility; interactions between the dimensions; leave no one behind; universality and accountability.*

- Where possible, the unintended effects and the possible potential and risks of interactions between the dimensions should be identified in multidisciplinary teams that include different sectoral perspectives. Should this lead to conflicting expectations, these can be documented in order to improve theory formation in the long term in a transparent and logical way. Highlighting such (possible) effects would facilitate evaluability, and would therefore also be conducive to the efficiency of an evaluation.
- The search for possible unintended effects could be supported by the use of existing frameworks such as the standards for environmental and social impact assessments, the IFC Performance Standards, the Environmental and Social Safeguards of the World Bank, the Environmental, Health and Safety Guidelines or the core labour standards of the International Labour Organization.
- Responses should also be found to the other principles of the 2030 Agenda, e.g. the mandate to 'leave no one behind'.

Recommendations on further developing the evaluation system

15. The evaluation team recommends that the BMZ develop an overarching evaluation strategy that in the course of time sets thematic priorities.
    - The implementing organisations might translate an overarching evaluation strategy of the BMZ into strategic evaluation programmes. The overarching strategy might also be supported by thematic cross-section evaluations conducted by DEval.
    - The design of the evaluation strategies and programmes to 2030 could be based on the principles of the 2030 Agenda and the accompanying system of goals and targets of the SDGs. The evaluation strategy might also be used to review the appropriateness of the degree of coverage by evaluations, and the preparation of appropriate sampling plans.
16. In the evaluation strategy the BMZ should define what requirements arise from the questions raised by the 2030 Agenda for the various evaluations – i.e. at the level of modules, programmes and country strategies.
    - The final assessment of the contributions made by German development cooperation to the SDGs and the principles of the 2030 Agenda could in the future take place chiefly at the level of programmes. Here it should be noted that there will continue to be many options in evaluations at the module level to assess the contribution to the 2030 Agenda's vision of sustainable development.
    - Since several actors are often involved at the programme level, the individual contributions could be captured and brought together in joint evaluations by the implementing organisations. Here it will also be necessary to clarify how to go about evaluating sector and global projects.
    - Selecting programme and module evaluations as part of the evaluation strategy could involve a two-stage selection process. A first step could be to select the programmes that would be suitable for evaluation. The decisions taken could also incorporate political deadlines (date of government negotiations, report of a partner country before the United Nations etc.). In a second step the module evaluations might then be selected.

6.

REFERENCES

Ashoff, G. (2015), Die Global Governance-Qualität der internationalen Aid Effectiveness Agenda: eine theoretische Analyse und Bewertung der Systemreform der internationalen Entwicklungszusammenarbeit, Deutsches Institut für Entwicklungspolitik, Bonn.

BMZ (2006), *Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit. Eine Orientierung für Evaluierungen des BMZ und der Durchführungsorganisationen,* Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung, Bonn/Berlin.

BMZ (2008), *Guidelines for bilateral Financial and Technical Cooperation with cooperation partners of German Development Cooperation,* No. 165, BMZ Konzepte, Bonn/Berlin.

Boone, P. (1996), *Politics and the effectiveness of foreign aid,* NBER Working Paper Series, Cambridge; Massachusetts, p. 289 – 329.

Carlsson, J. and L. Wohlgemuth (1996), *Capacity Building and Networking - A meta-evaluation of African regional research networks,* Sida Evaluation, Department for Evaluation and Internal Audit, Stockholm.

Caspari, A. (2004), Evaluation der Nachhaltigkeit von Entwicklungszusammenarbeit. Zur Notwendigkeit angemessener Konzepte und Methoden, Sozialwissenschaftliche Evaluationsforschung, VS Verlag für Sozialwissenschaften, Wiesbaden.

Cutter, A. (2014), *Sustainable Development Goals (SDG) and Integration: Achieving a better balance between the economic, social and environmental dimensions,* German Council for Sustainable Development, Berlin.

Dietz, F. and A. Hanemaaijer (2012), How to select policy-relevant indicators for sustainable development, in von Raggamby, A. and F. Rubik (eds.), Sustainable development, evaluation and policy-making: theory, practise and quality assurance, Edward Elgar, Cheltenham, p. 21 – 35.

Faust, J. (2007), Assessing Aid: die makroquantitative Forschung zur Effektivität der Entwicklungszusammenarbeit, in Hemmer, H.-R. (ed.), *Zur Wirksamkeitsdebatte in der Entwicklungszusammenarbeit (EZ),* Erfurt.

Freedom House (2016), *Freedom in the World,* New York.

GIZ (2016), *Meta-Evaluierung der Projektevaluierungen (PEV),* Bonn.

Grunwald, A. und J. Kopfmüller (2006), Nachhaltigkeit, Campus Einführungen, Campus-Verlag, Frankfurt am Main.

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

Hartmuth, G. (2004), *Nachhaltige Entwicklung im lokalen Kontext – Schritte zur Entwicklung eines kommunalen Nachhaltigkeits-Indikatorensystems,* No. 6, UFZ Diskussionspapiere, Umweltforschungszentrum, Leipzig.

Islam, S.M.N. and M.F. Clarke (2005), The welfare economics of measuring sustainability: a new approach based on social choice theory and systems analysis, Sustainable Development, Vol. 13, No. 5, p. 282 – 296.

KfW Entwicklungsbank (2003), *FZ-Projekte und Nachhaltigkeit. Zur Berücksichtigung der Nachhaltigkeit durch die KfW in Schlussprüfungen von FZ-Vorhaben: Grundsätzliche Überlegungen, No. 33,* Diskussionsbeiträge, KfW Entwicklungsbank, Frankfurt am Main.

KfW Entwicklungsbank (2016), *KfW Nachhaltigkeitsrichtlinie,* KfW Entwicklungsbank, Frankfurt am Main.

Klasen, S. (2015), *SDG – Den Ärmsten der Welt einen Bärendienst erwiesen,* No. 3, Meinungsforum Entwicklungspolitik, KfW Entwicklungsbank, Frankfurt am Main.

Klingebiel, S. (2013), Entwicklungszusammenarbeit – eine Einführung, Deutsches Institut für Entwicklungspolitik, Bonn.

König, J. and J. Thema (eds.) (2011), Nachhaltigkeit in der Entwicklungszusammenarbeit: theoretische Konzepte, strukturelle Herausforderungen und praktische Umsetzung, Globale Gesellschaft und internationale Beziehungen, Verlag für Sozialwissenschaft, Wiesbaden, 1. Auflage.

Landis, J.R. and G.G. Koch (1977), The Measurement of Observer Agreement for Categorical Data, *Biometrics,* Vol. 33, No. 1, p. 159.

Leeuw, F.L. and L.J. Cooksy (2005), Evaluating the performance of development agencies: The role of metaevaluations., in Pitman, G.K., O.N. Feinstein and G.K. Ingram (eds.), *Evaluating development effectiveness,* World Bank series on Evaluation and Development, Transaction, New Brunswick, p. 95 – 108.

Meadows, D.H. et al. (1972), The limits to growth: a report for the Club of Rome's Project on the Predicament of Mankind, Universe Books, New York.

Noltze, M. et al. (2018), *Evaluation synthesis on sustainability in German development cooperation,* German Institute for Development Evaluation (DEval), Bonn.

Nuscheler, F. (2007), Wie geht es weiter mit der Entwicklungspolitik?, *Aus Politik und Zeitgeschichte,* Vol. 48, p. 3 – 10.

OECD (1991), *DAC Criteria for Evaluating Development Assistance Factsheet,* Paris.

OECD (2008), *The Paris Declaration on Aid Effectiveness and the Accra Agenda for Action,* Paris.

OECD (2010a), Evaluation in Development Agencies, Better Aid, OECD Publishing, Paris.

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

OECD (2016a), Better Policies for Sustainable Development 2016. A New Framework for Policy Coherence, OECD Publishing, Paris.

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

OECD (2017), DAC Criteria for Evaluating Development Assistance, *Organisation for Economic Co-operation and Development, http://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm,* acceses on 23/03/2017.

Patton, M.Q. (2008), Utilization-focused evaluation, SAGE Publications, Thousand Oks, Calif., 4. Aufl.

Preiß, J. (2017), *Evaluierung von Nachhaltigkeit und ihre Determinanten in der Entwicklungszusammenarbeit. Eine empirische Analyse anhand von Projekten der Weltbank,* unveröffentlichte Masterarbeit, Freie Universität Berlin, Berlin.

Raetzell, L. and M. Krämer (2013), *Meta-Evaluation Gesundheit,* Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), Bonn.

Russ-Eft, D.F. (2014), Human resource development, evaluation, and sustainability: what are the relationships?, *Human Resource Development International,* Vol. 17, No. 5, p. 545 – 559.

Scriven, M. (1991), Evaluation Thesaurus, Sage Publications, Newbury Park; London; New Delhi, 4th edition

Scriven, M. (2009), Meta-Evaluation revisited, *Journal of MultiDisciplinary Evaluation,* Vol. 6, No. 11, p. iii – viii.

Stadtler, L. (2016), Scrutinizing Public–Private Partnerships for Development: Towards a Broad Evaluation Conception, *Journal of Business Ethics,* Vol. 135, No. 1, p. 71 – 86.

Stockmann, R. and W. Gaebe (eds.) (1993), Hilft die Entwicklungshilfe langfristig? Bestandsaufnahme zur Nachhaltigkeit von Entwicklungsprojekten, Westdeutscher Verlag GmbH, Opladen.

Stockmann, R. and S. Silvestrini (2011), *Synthese und Metaevaluierung Berufliche Bildung,* Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn/Eschborn.

Stockmann, R. and S. Silvestrini (2012), Ergebnispräsentation: Synthese und Meta-Evaluierung Berufliche Bildung, gehalten auf der GIZ Dialogtag, Bonn.

Stufflebeam, D.L. (2001), The Metaevaluation Imperative, American Journal of Evaluation, Vol. 22, No. 2, p. 183 – 209.

van Tulder, R. and S. Pfisterer (2008), *From Idea to Partnership: Reviewing the Effectiveness of Development Partnerships in Zambia, Columbia and Ghana,* Expert Centre for Sustainable Business & Development Cooperation, Maastricht.

UN (2015), Transforming our world. The 2030 Agenda for Sustainable Development, New York.

Vahlhaus, M. (2014), *Der Weg: Scaling-Up. Das Ziel: Breitenwirksamkeit,* Deutsche Gesellschaft für Internationale Zusammenarbeit, Bonn, Eschborn.

Von Raggamby, A. and F. Rubik (Hrsg.) (2012), Sustainable development, evaluation and policy-making: theory, practise and quality assurance, Evaluating sustainable development, Edward Elgar, Cheltenham.

Widmer, T. (2006), Meta-Evaluation. Kriterien zur Bewertung von Evaluationen, Verlag Paul Haupt, Zürich.

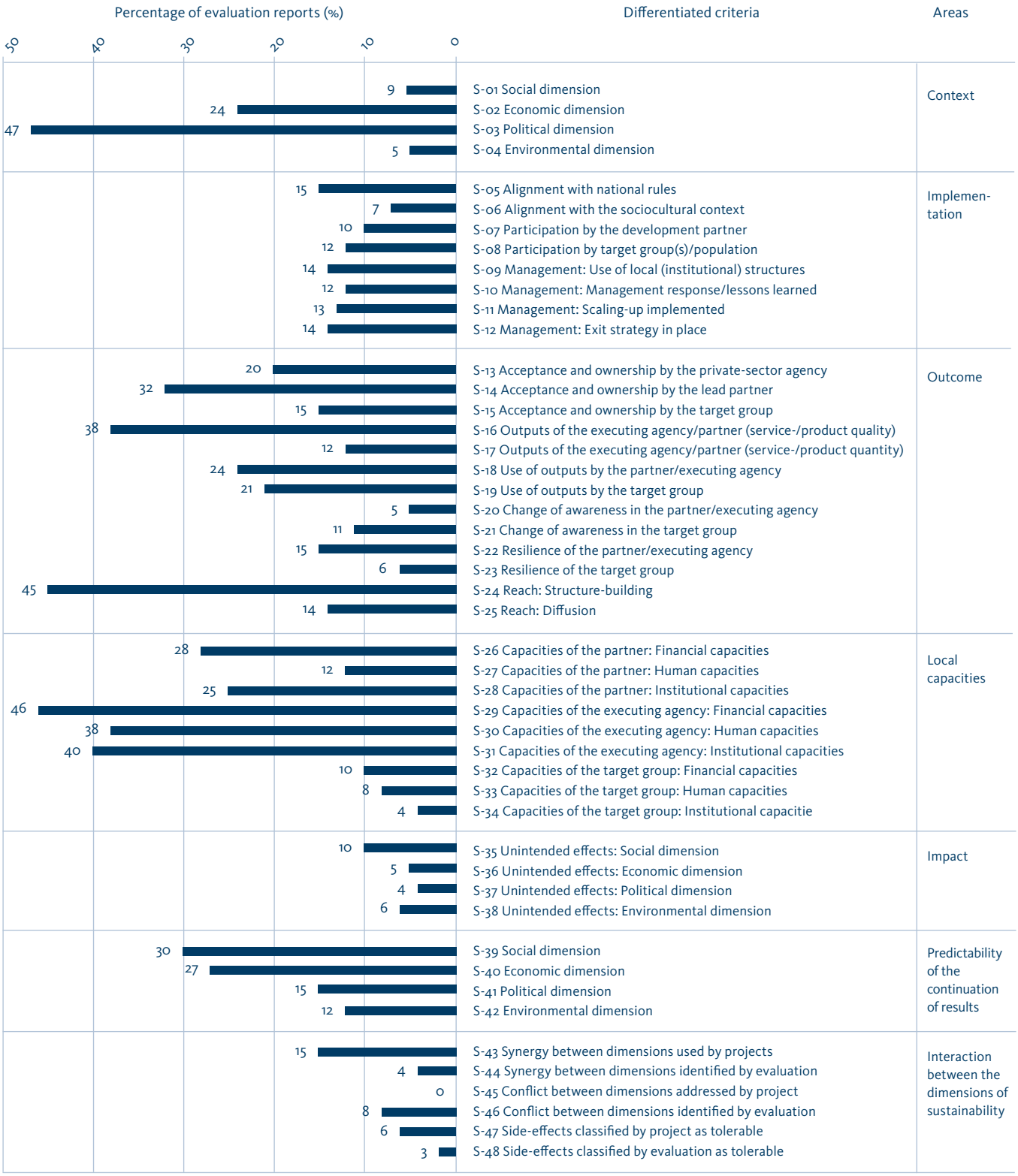
World Commission on Environment and Development (1987), Our Common Future, Oxford University Press, Oxford.

7.

ANNEX

7.1 Figures

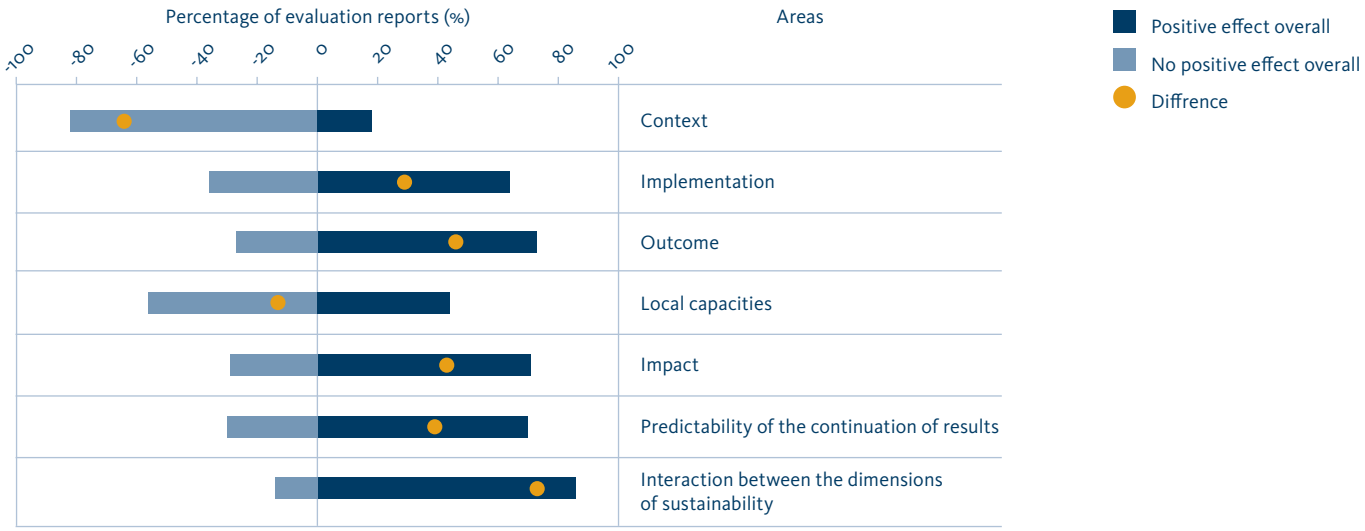
Figure 11: Percentage of evaluation reports referring to differentiated sustainability criteria



Source: authors' own graphic.

Notes: The graphic shows the percentage of evaluation reports that refer to the relevant differentiated criterion when assessing sustainability. N = 513.

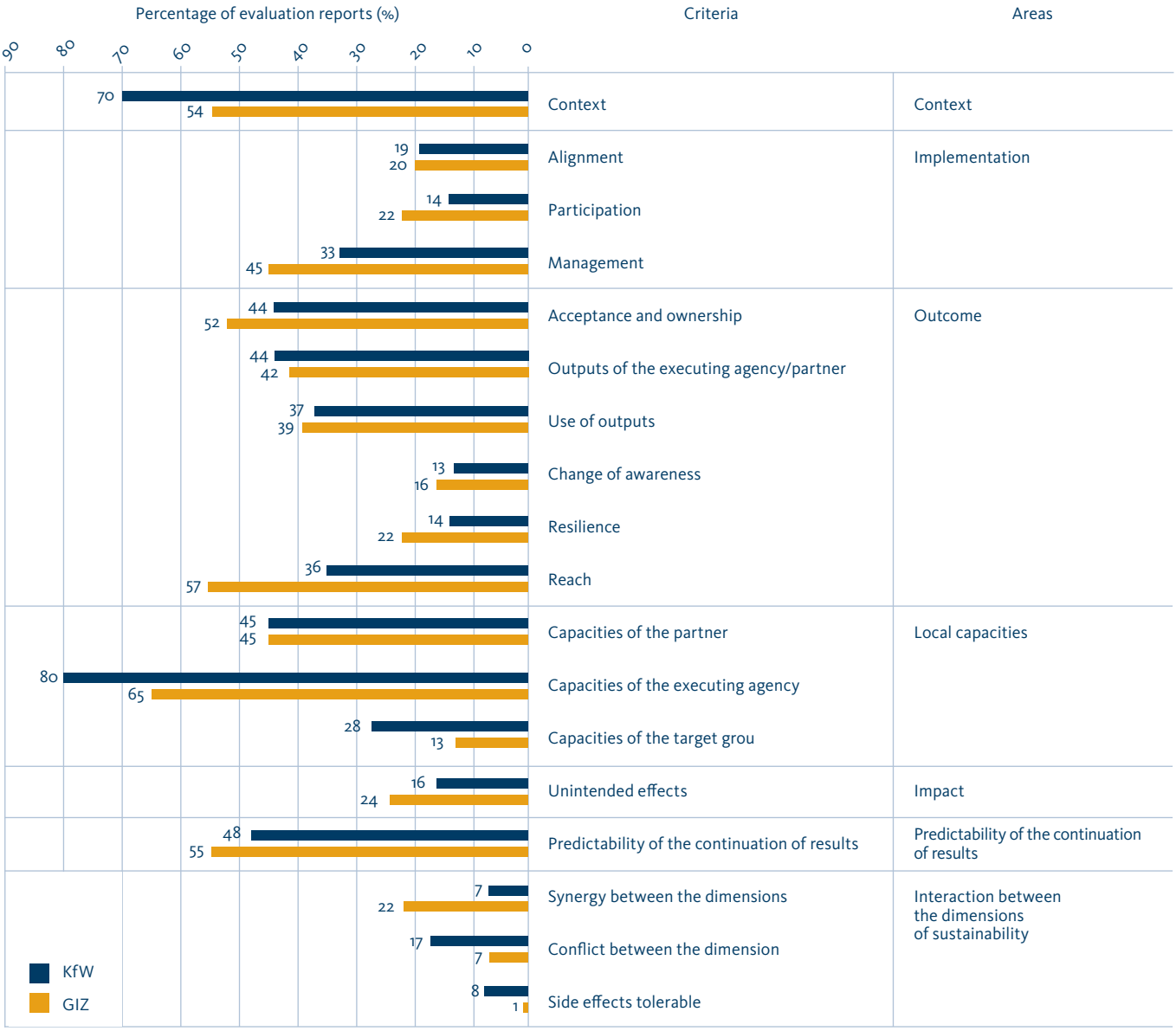
Figure 12: Percentage of evaluation reports by differentiated sustainability area and effect on sustainability assessment



Source: authors' own graphic.

Notes: The bars show the percentage of evaluation reports that ascribe to each area a positive or negative effect on the sustainability of a project. Individual areas include only those reports that refer to at least one differentiated criterion per area covered when assessing sustainability. The dots represent the difference between the percentages of positive and negative assessments of an area. N = 513.

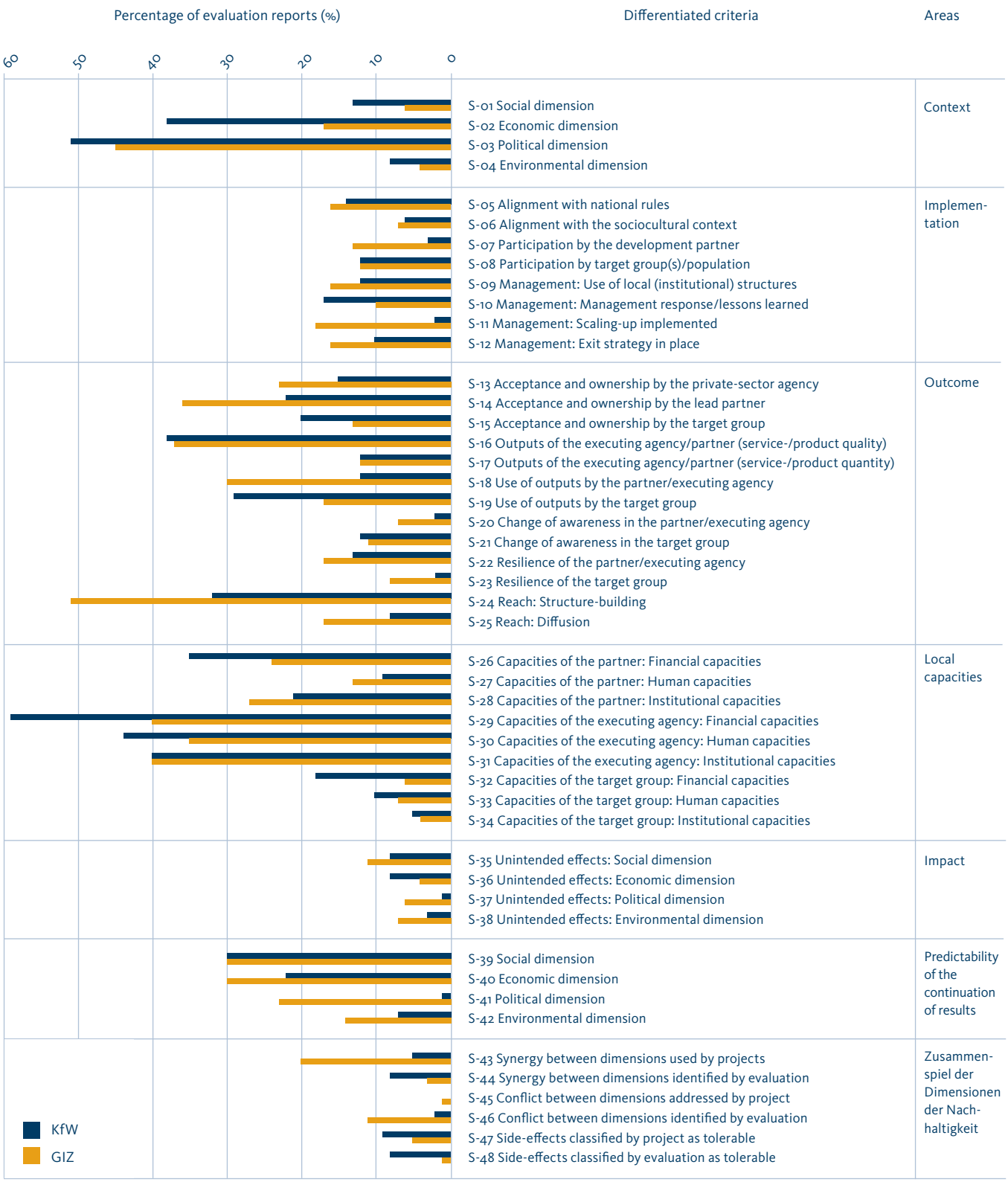
Figure 13: Percentage of evaluation reports referring to sustainability criteria by implementing organisation



Source: authors' own graphic

Notes: The graphic shows the percentage of evaluation reports that refer to at least one differentiated criterion for the respective sustainability criterion when assessing sustainability. The evaluation reports are broken down into KfW (n = 172) and GIZ (n = 341). N = 513.

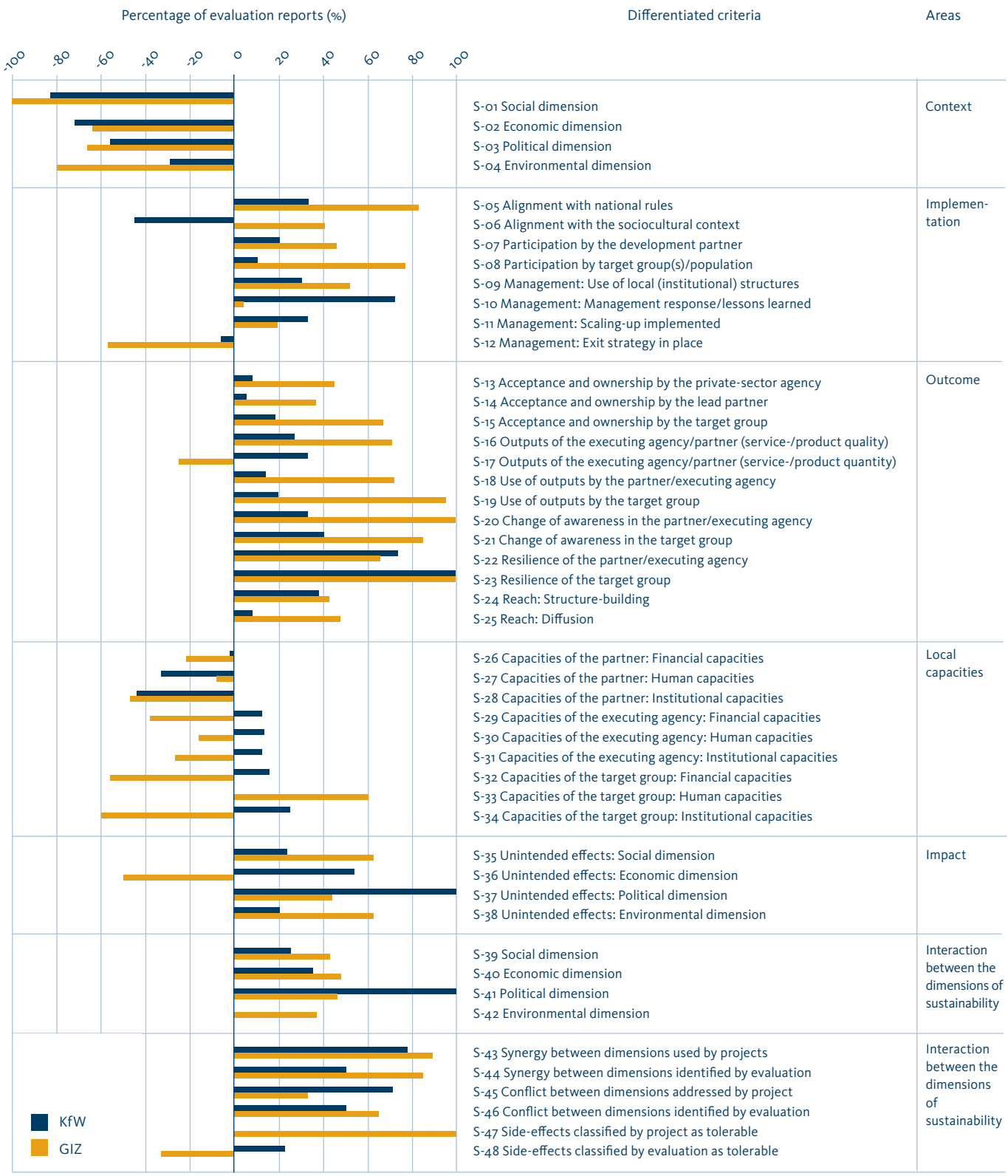
Figure 14: Percentage of evaluation reports referring to differentiated sustainability criteria by implementing organisation



Source: authors' own graphic.

Notes: The graphic shows the percentage of evaluation reports that refer to the relevant differentiated criteria when assessing sustainability. The evaluation reports are broken down into KfW (blue, n = 172) and GIZ (orange, n = 341). N = 513.

Figure 15: Percentage of evaluation reports referring to differentiated sustainability criteria and effect on sustainability assessment by implementing organisation



Source: authors' own graphic.

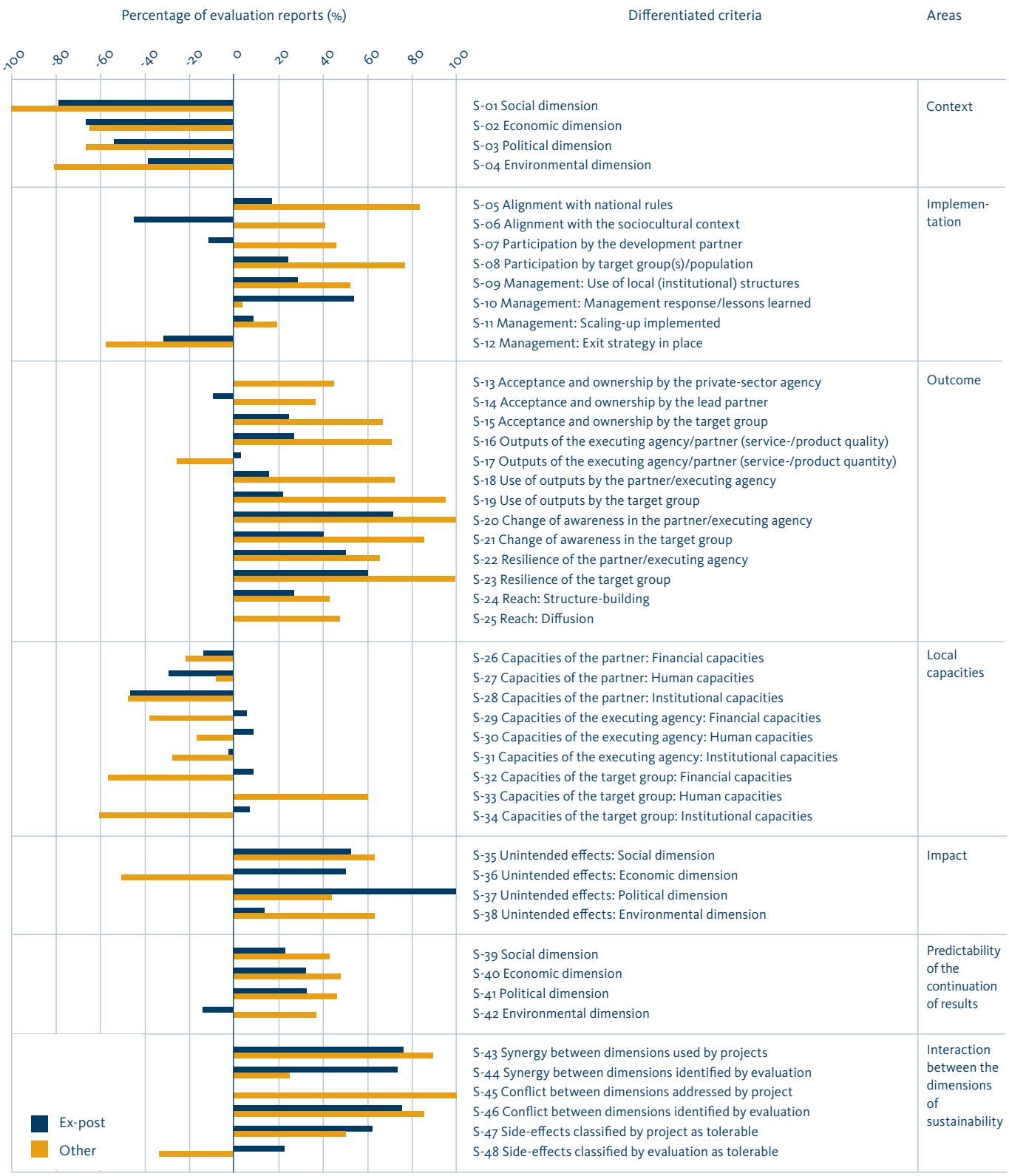
Notes: The bars show the percentage of evaluation reports that ascribe to the respective differentiated criterion either a positive or a negative effect on the sustainability of a project. The evaluation reports are broken down into KfW (n = 172) and GIZ (n = 341). N = 513.

Figure 16: Percentage of evaluation reports referring to sustainability criteria and effect on sustainability assessment by implementing organisation

Source: authors' own graphic.

Notes: The bars show the percentage of evaluation reports that ascribe to the respective criterion either a positive or a negative effect on the sustainability of a project. The evaluation reports are broken down into KfW (n = 172) and GIZ (n = 341). N = 513.

Figure 17: Percentage of evaluation reports referring to differentiated sustainability criteria and effect on sustainability assessment by evaluation type



Source: authors' own graphic.

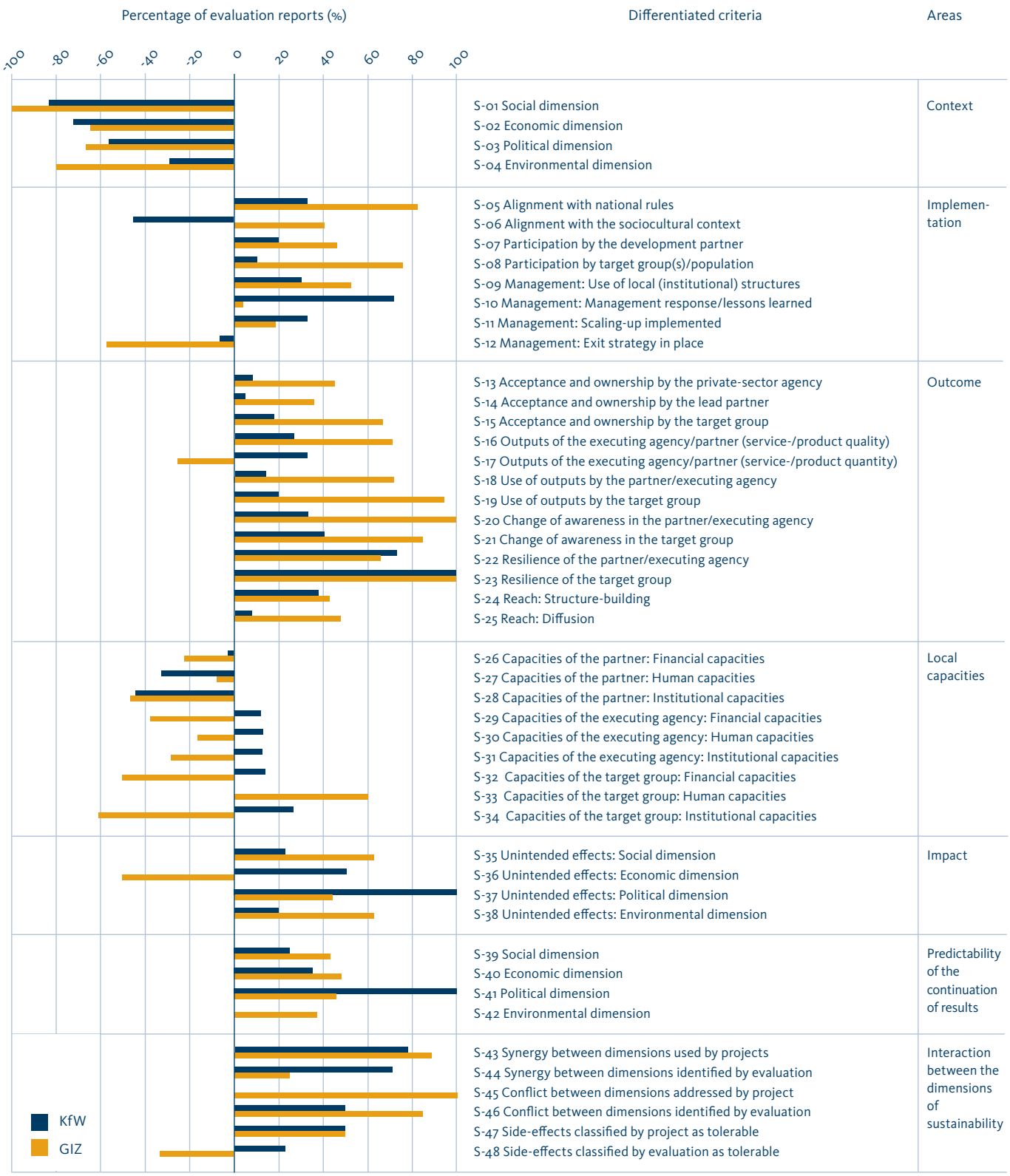
Notes: The bars show the percentage of evaluation reports that ascribe to the respective differentiated criterion either a positive or a negative effect on the sustainability of a project. The evaluation reports are broken down into ex-post-evaluations (blue, n = 219) and PPRs. PEs and final evaluations (orange, n = 294). N = 513.

Figure 18: Percentage of evaluation reports referring to sustainability criteria and effect on sustainability assessment by evaluation type

Source: authors' own graphic.

Notes: The bars show the percentage of evaluation reports that ascribe to the respective sustainability criterion a positive or negative effect on the sustainability of a project. The evaluation reports are broken down into ex-post-evaluations (blue, n = 219) and PPRs. PEs and final evaluations (orange, n = 294). N = 513.

Figure 19: Percentage of ex-post evaluation reports referring to differentiated sustainability criteria and effect on sustainability assessment by implementing organisation



Source: authors' own graphic.

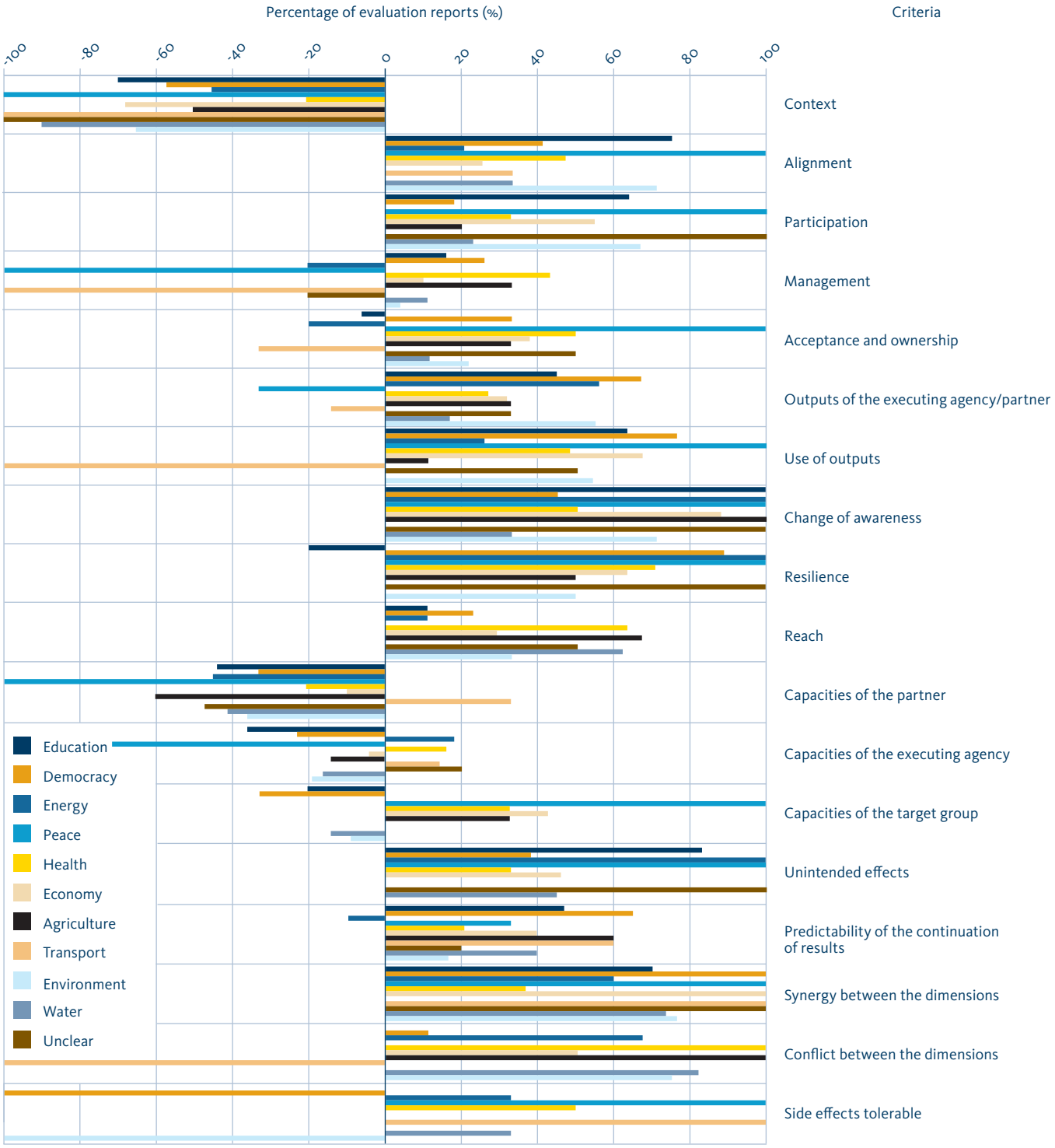
Notes: The bars show the percentage of evaluation reports that ascribe to the respective differentiated criterion either a positive or a negative effect on the sustainability of a project. The graphic shows only ex-post evaluations. These are broken down into KfW (blue, n = 172) and GIZ (orange, n = 38). N = 210.

Figure 20: Percentage of ex-post evaluations referring to sustainability criteria and effect on sustainability assessment by implementing organisation

Source: authors' own graphic.

Notes: The bars show the percentage of evaluation reports that ascribe to the respective sustainability criterion a positive or negative effect on the sustainability of a project. The graphic shows only ex-post evaluations. These are broken down into KfW (blue, n = 172) and GIZ (orange, n = 38). N = 210.

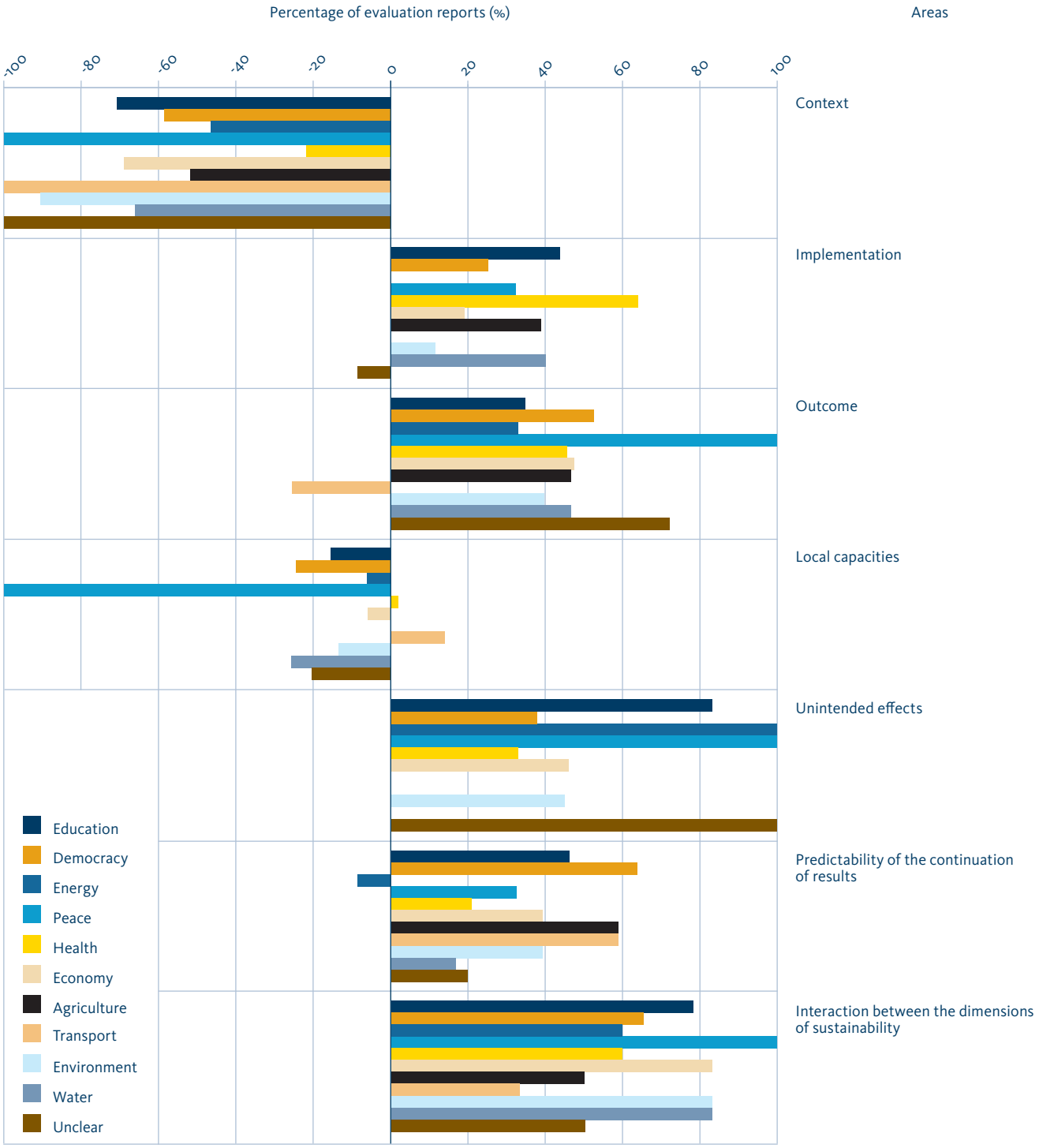
Figure 21: Percentage of evaluation reports referring to sustainability criteria and effect on sustainability assessment by sector



Source: authors' own graphic.

Notes: The bars show the percentage of evaluation reports that ascribe to the respective sustainability criterion a positive or negative effect on the sustainability of a project. The evaluation reports are broken down by the sector in which the project is implemented. N = 513.

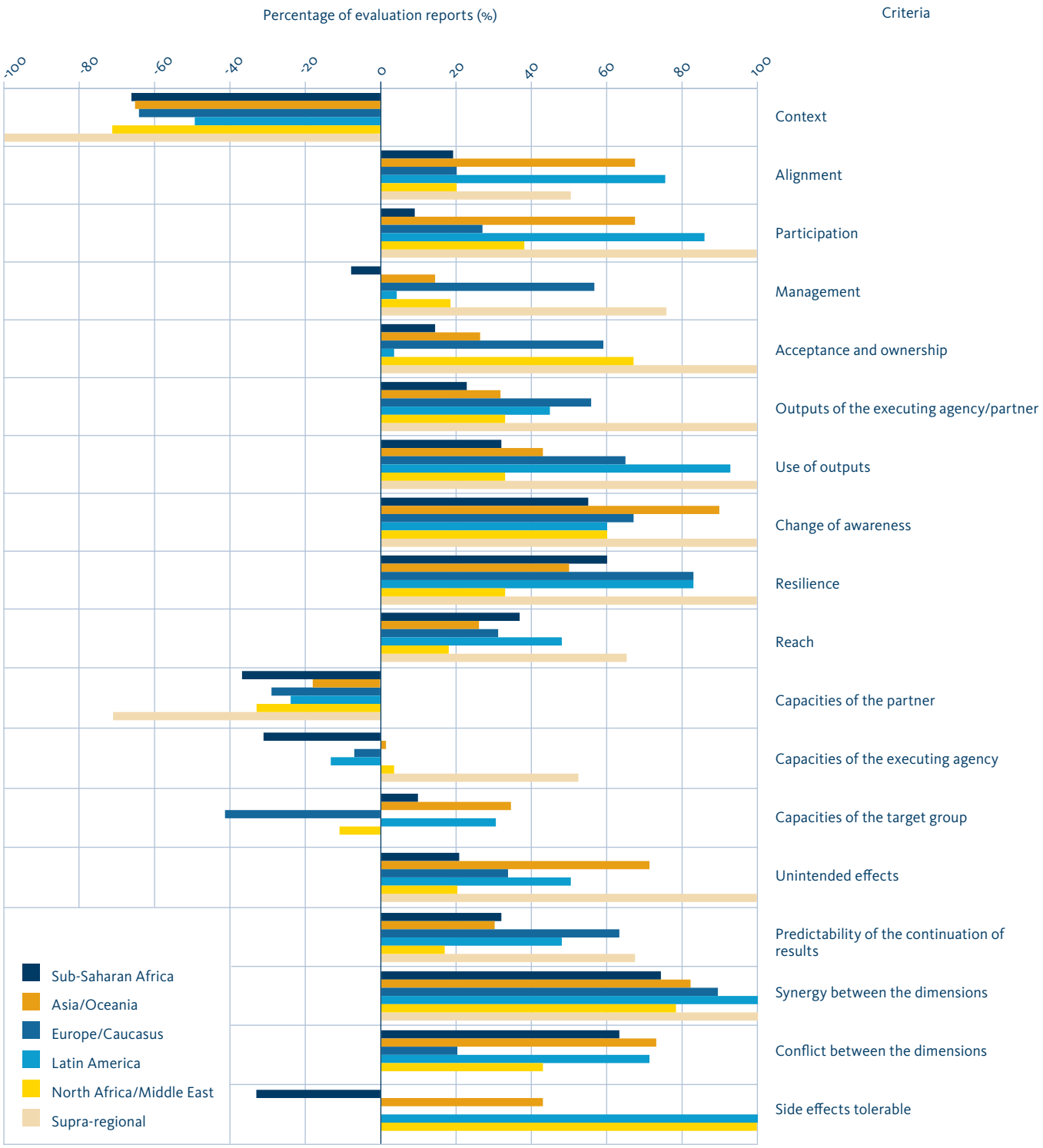
Figure 22: Percentage of evaluation reports referring to sustainability areas and effect on sustainability assessment by sector



Source: authors' own graphic.

Notes: The bars show the percentage of evaluation reports that ascribe to the respective sustainability criterion a positive or negative effect on the sustainability of a project. The evaluation reports are broken down by the sector in which the project is implemented. N = 513.

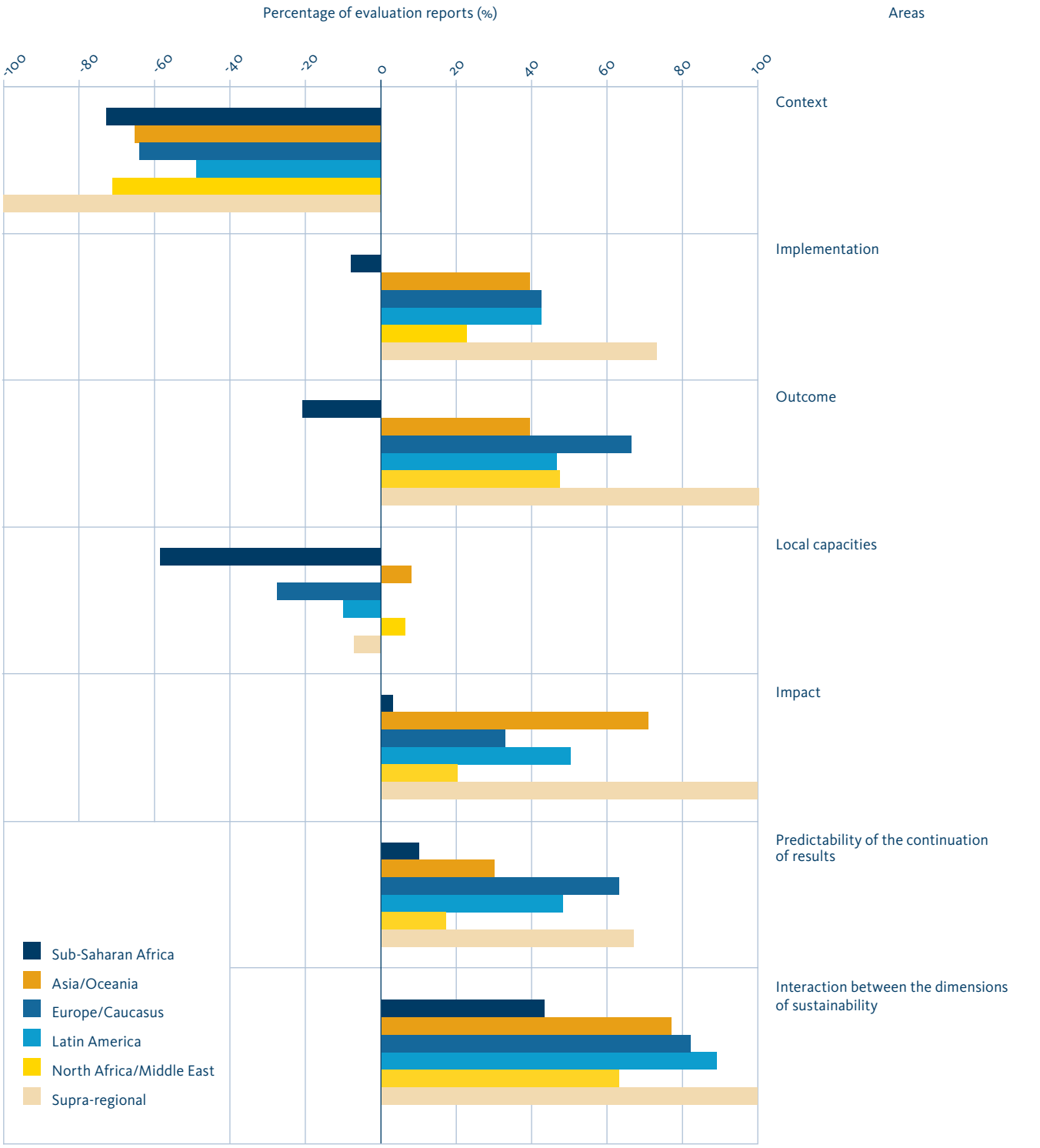
Figure 23: Percentage of evaluation reports referring to sustainability criteria and effect on sustainability assessment by region



Source: authors' own graphic.

Notes: The bars show the percentage of evaluation reports that ascribe to the respective sustainability criteria a positive or negative effect on the sustainability of a project. The evaluation reports are broke down by the region in which the project is implemented. N = 513.

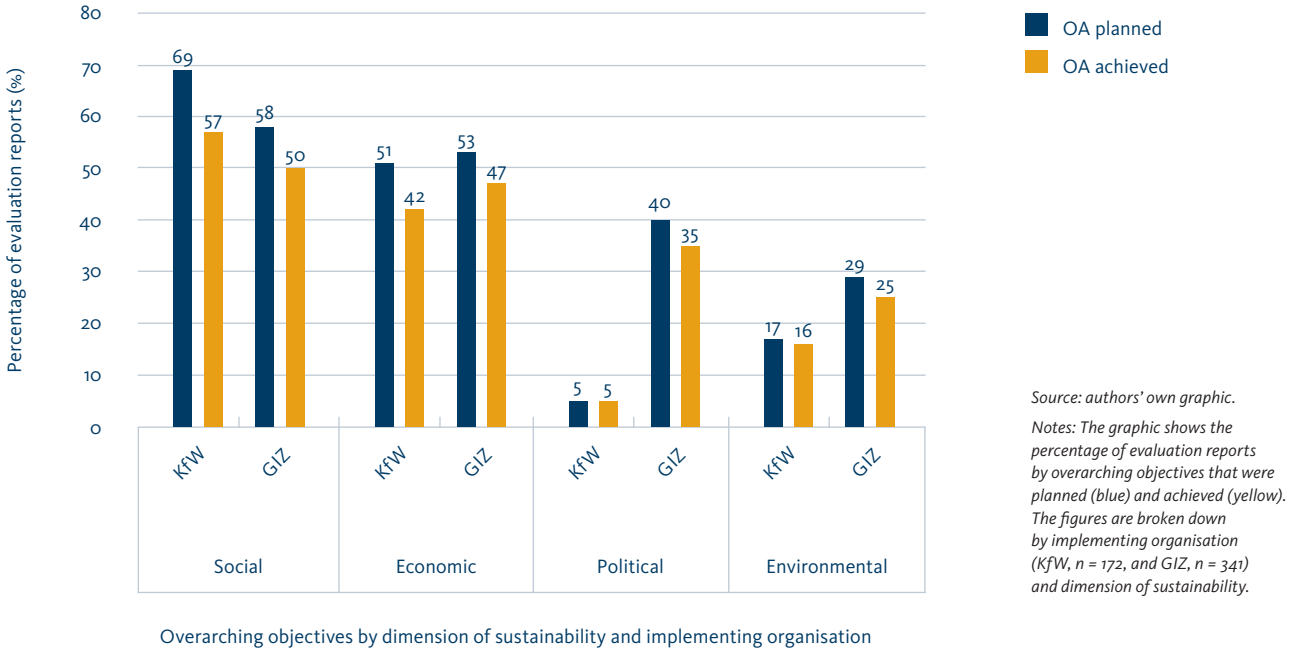
Figure 24: Percentage of evaluation reports referring to sustainability areas and effect on sustainability assessment by region



Source: authors' own graphic.

Notes: The bars show the percentage of evaluation reports that ascribe to the respective sustainability criterion a positive or negative effect on the sustainability of a project. The evaluation reports are broken down by the sector in which the project is implemented. N = 513.

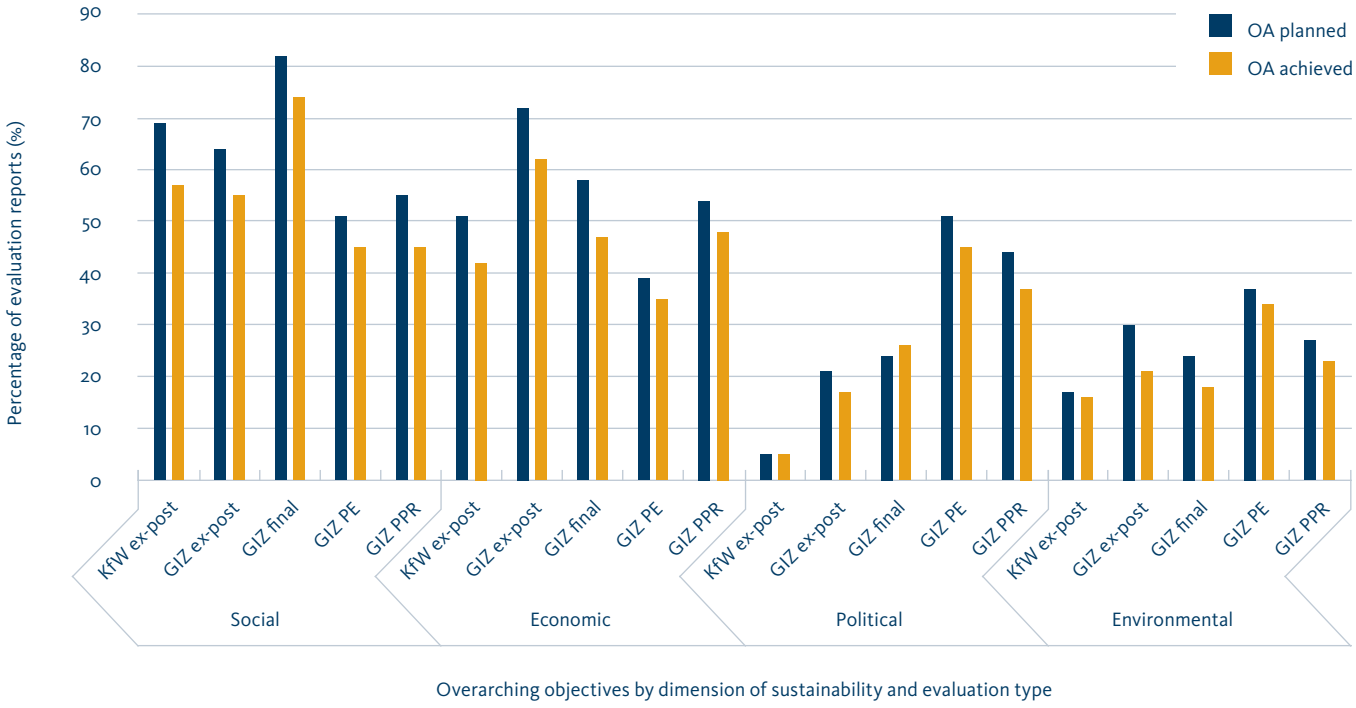
Figure 25: Percentage of evaluation reports by planned and achieved overarching objectives by implementing organisation



Source: authors' own graphic.

Notes: The graphic shows the percentage of evaluation reports by overarching objectives that were planned (blue) and achieved (yellow). The figures are broken down by implementing organisation (KfW, n = 172, and GIZ, n = 341) and dimension of sustainability.

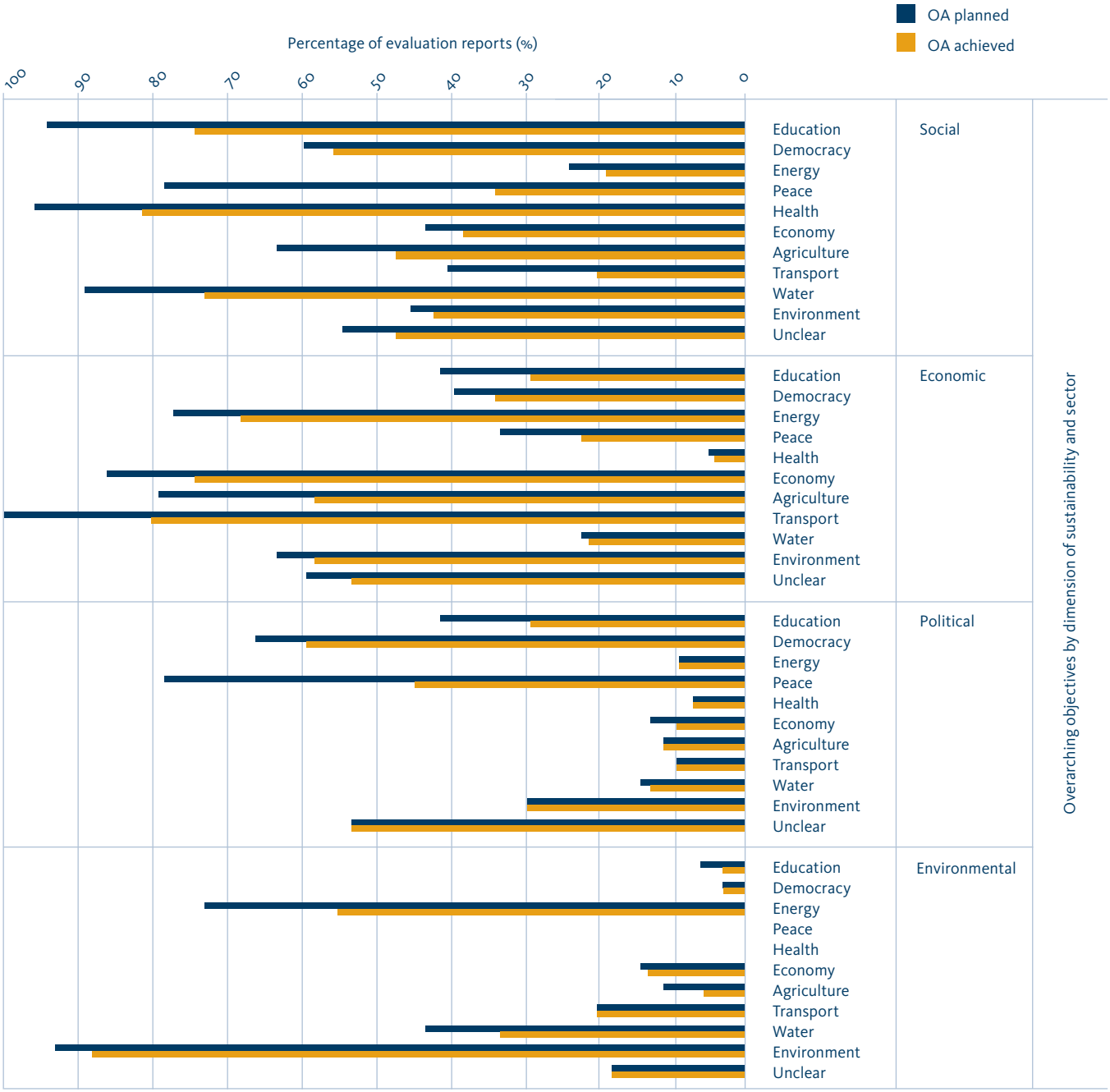
Figure 26: Percentage of evaluation reports by planned and achieved overarching objective, evaluation type and sustainability dimension



Source: authors' own graphic.

Notes: The graphic shows the percentage of evaluation reports by overarching objective and achievement of the overarching objective. The evaluation reports are broken down by evaluation type. These are: KfW-ex-post-evaluations (n = 172), GIZ-ex-post evaluations (n = 47), GIZ-final evaluations (n = 38), GIZ PEs (n=82), and GIZ PPRs (n = 174). Within a pair of bars for an evaluation type, the overarching objectives of the project are broken down by overarching objectives planned and those actually achieved.

Figure 27: Percentage of evaluation reports by planned and achieved overarching objective, sector and sustainability dimension



Source: authors' own graphic.

Notes: The graphic shows the percentage of evaluation reports by overarching objective and achievement of the overarching objective. The evaluation reports are broken down by the sector in which the project is implemented. These are: education (n = 34), democracy (n = 95), energy (n = 22), peace (n = 9), health (n = 57), economy (n = 127), agriculture (n = 19), transport (n = 10), water (n = 63), environment (n = 60), and 'not clear' (n = 17). Within a pair of bars for a sector, the overarching objectives of the project are broken down by overarching objectives planned and those actually achieved.

7.2 Tables

Table 4: Analysis grid for the assessment of evaluation quality

Areas	No.[22]	Criteria	Definition of the criterion
1. Evaluation background	Q-01	Object (project) described	The criterion is met when the 1) objectives, 2) target group, 3) context and 4) relevant actors (partner and / or executing agency) of the development cooperation project are described and the object has thus been defined.
	Q-02	Area of enquiry formulated / operationalised	The criterion is met when the area of enquiry and / or evaluation questions are specified / concretised.
2. Description of the causal relationships	Q-03	Results logic / results chain described	The criterion is met when the description of the intended results of the development cooperation project distinguishes between different levels of results (input-output-outcome-impact), and these levels are linked through a logical sequence (and / or result hypotheses are formulated).
	Q-04	Results logic largely operationalised through indicators	The criterion is met when the degree to which objectives have been achieved is made measurable / is assessed using indicators, for the majority of programme objectives.
3. Methodology	Q-05	Methodology described	The criterion is met when the steps of the procedure for collecting and analysing data that will be used in the evaluation are described and operationalised.
	Q-06	Strengths and / or limitations of the methodology identified	The criterion is met when a rationale is in place to explain why the methods applied are appropriate to the object of the evaluation. Advantages and limitations of the methodology are discussed.
	Q-07	Respondents identified	The criterion is met when the persons to be consulted / surveyed in order to collect data have been identified.
	Q-08	Selection procedure for respondents described	The criterion is met when the selection of persons to be consulted / surveyed and selection criteria have been described.
4. Data collection methods		Analysis of documents / databases	The criterion is met when documents and / or data from secondary databases are analysed.
		Monitoring data used	The criterion is met when monitoring data are analysed.
		Semi-structured interviews	The criterion is met when semi-structured interviews are used.
		Standardised interviews	The criterion is met when standardised interviews are used.
		Focus group discussion	The criterion is met when focus group discussions are used.
		Participatory methods	The criterion is met when participatory data collection methods (problem tree, SWOT analysis etc.) are used and/or the participants help develop the topics to be discussed.
		Systematic observations	The criterion is met when systematic observations (on-site inspections, sample testing) are performed.

[22] *A number 'Q-.....' is assigned to all those criteria included in the assessment as part of the quality index due to their explanatory significance regarding the quality of the evaluation reports.*

5. Evaluation design	Q-09	Before and after comparison	The criterion is met when the results of the development cooperation programme are determined by comparing values for the majority of all indicators at the beginning of the project with values after the project has come to an end.
	Q-10	Control / comparison group included	The criterion is met when the outcomes of an intervention group (within the sphere of influence of the development cooperation project) are compared to the outcomes of a control group (beyond the sphere of influence of the development cooperation project).
	Q-11	Causality inferred on the basis of plausibility	The criterion is met when the results of the development cooperation project are inferred using a systematic procedure based on plausibility (especially theory-based approaches, e. g. contribution analysis).
6. Robustness of the findings	Q-12	Data triangulation applied	The criterion is met when the data on which the analysis is based originate from various sources (meaning various stakeholder groups and/or data collection tools) (> 1 source).
	Q-13	Triangulation methods applied	The criterion is met when data from the same source is analysed using various methods (> 1 method).
		Investigator triangulation	The criterion is met when at least two investigators are involved in the analysis, and when the report makes clear in its conclusions which investigator(s) support(s) this conclusion and which do(es) not.[23]
7. Analysis and conclusions	Q-14	Conclusions largely referenced through data	The criterion is met when the vast majority of findings and conclusions are placed in relation to the database analysis.
	Q-15	Conclusions from data largely plausibly substantiated	The criterion is met when the vast majority of findings and conclusions concerning results are made plausible on the basis of the data used.
	Q-16	Database sufficient with respect to conclusions	The criterion is met when the database and the methodology are qualitatively and quantitatively sufficient to draw the conclusions expressed (regarding results achieved).

Source: Authors' own table

[23] *Due to the practical difficulties associated with applying investigator triangulation in evaluation reports, no further use was made of this criterion in the analysis.*

Table 5: Analysis grid for the assessment of sustainability

Areas	Criteria	No.	Differentiated criteria	Definition
1) Context	1. Context by dimension	S-01	Social dimension	The criterion is met when the reported contextual factors have a direct effect on a) the results of the project or b) the predictability of the continuation of its results.
		S-02	Economic dimension	
		S-03	Political dimension	
		S-04	Environmental dimension	
2) Implementation	2. Alignment	S-05	Alignment with national rules	The criterion is met when the project coincides with a national strategy / a national programme.
		S-06	Alignment with the sociocultural context at the level of target groups	The criterion is met when the project coincides with social conventions.
	3. Participation	S-07	Participation by the development partner	The criterion is met when the executing agency / partner was at least consulted on decisions concerning implementation.
		S-08	Participation by target group(s) / population	The criterion is met when the target group(s) was / were at least consulted on decisions concerning implementation.
	4. Management	S-09	Use of local (institutional) structures	The criterion is met when existing official bodies, working groups or other institutional structures in the partner country or region are used to implement the project
		S-10	Management response / learning from monitoring and evaluation / lessons learned	The criterion is met when monitoring / evaluation results have been considered in project structures and / or project processes.
		S-11	Scaling-up strategy	The criterion is met when the activities have been extended to one or more provinces and / or target groups / stakeholder groups, and / or pilot projects have been systematised – e. g. – when several programme lines have been completed and transferred into larger programmes / a national strategy.
		S-12	Exit strategy	The criterion is met when a strategy for continuing the activities without German development cooperation was jointly developed with the partner / executing agency and / or steps have been described for gradually reducing the inputs or continuing the activity of German development cooperation after the end of the project on a reduced basis.

Areas	Criteria	No.	Differentiated criteria	Definition
3) Outcome	5. Acceptance and ownership	S-13	Acceptance and ownership by the private-sector agency	The criterion is met when the private-sector agency has shown initiative and / or very largely kept pledges / discharged its own obligations and / or assumed responsibility.
		S-14	Acceptance and ownership by the partner	The criterion is met when the private-sector agency has shown initiative and / or very largely kept pledges / discharged its own obligations and / or assumed responsibility.
		S-15	Acceptance and ownership by the target group.	The criterion is met when the target group has shown initiative and /or very largely kept pledges/discharged its own obligations and / or assumed responsibility.
	6. Outputs of the executing agency / partner	S-16	Service / product quality	The criterion is met when the quality of the output is assessed as largely sufficient for achieving the programme objectives.
		S-17	Service / product quantity	The criterion is met when the quantity of the output is assessed as largely sufficient for achieving the programme objectives.
	7. Use of outputs	S-18	Use of outputs by the partner / executing agency	The criterion is met when project outputs (strategies, materials) are being used by the partner / executing agency.
		S-19	Use of outputs by the target group	The criterion is met when project outputs (strategies, materials) are being used by the target group.
	8. Change of awareness	S-20	Change of awareness in the partner / executing agency	The criterion is met when the partner / executing agency is seen to have undergone a change of awareness beyond the use of outputs (manifested by changes in behaviour also outside the project / without incentives).
		S-21	Change of awareness in the target group	The criterion is met when the target group is seen to have undergone a change of awareness beyond the use of outputs (manifested by changes in behaviour also outside the project / without incentives).
	9. Resilience and adaptability	S-22	Resilience and adaptability of the partner / executing agency	The criterion is met when the partner / executing agency is able to recognise chances and opportunities for themselves and act accordingly.
		S-23	Resilience and adaptability of the target group	The criterion is met when the target group is able to recognise chances and opportunities for itself and act accordingly.
	10. Reach	S-24	Structure-building (direct)	The criterion is met when changes take place not only at the level of individuals but also at the level of systems.
		S-25	Diffusion (indirect)	The criterion is met when concepts or ideas are transferred to people who were not part of the original target group.

Areas	Criteria	No.	Differentiated criteria	Definition
4) Local capacities	11. Capacities of the partner	S-26	Financial / economic inputs	The criterion is met when financial/economic inputs to be provided by the partner are provided as agreed/when the inputs are sufficient for successful continuation of the activities.
		S-27	Human capacities / expertise	The criterion is met when a) sufficient personnel are available and b) the personnel are sufficiently well qualified to successfully continue the project activities.
		S-28	Institutional / organisational inputs	The criterion is met when a sufficient degree of institutional independence and organisational effectiveness/efficiency is in place in order to achieve programme objectives/when institutional inputs are provided as agreed.
	12. Capacities of the executing agency	S-29	Financial / economic inputs	The criterion is met when financial / economic inputs to be provided by the executing agency are provided as agreed/when the inputs are sufficient for successful continuation of the activities.
		S-30	Human capacities / expertise	The criterion is met when a) sufficient personnel are available and b) the personnel are sufficiently well qualified to successfully continue the project activities.
		S-31	Institutional / organisational capacities	The criterion is met when a sufficient degree of institutional independence and organisational effectiveness / efficiency is in place in order to achieve programme objectives.
	13. Capacities of the target group	S-32	Financial / economic inputs	The criterion is met when financial / economic inputs to be provided by the target group are provided as agreed/when the inputs are sufficient for successful continuation of the activities.
		S-33	Human capacities / expertise	The criterion is met when the targets groups are sufficiently well qualified / procurement of the needed expertise is guaranteed, such that the project activities can be continued successfully.
		S-34	Institutional / organisational capacities	The criterion is met when a sufficient degree of institutional independence and organisational effectiveness/efficiency to achieve programme objectives is in place on the part of the user.
5) Impact	14. Unintended effects by dimension	S-35	Social aspects	The criterion is met when the project leads to changes outside of the overarching objective / programme objective.
		S-36	Economic aspects	
		S-37	Political aspects	
		S-38	Environmental aspects	
6) Predictability of the continuation of results	15. Predictability of the continuation of results by dimension	S-39	Social aspects	The criterion is met when the factors that safeguard continuation of the positive results or increase the results predominate.
		S-40	Economic aspects	
		S-41	Political aspects	
		S-42	Environmental aspects	

Areas	Criteria	No.	Differentiated criteria	Definition
7) Interaction between the dimensions of sustainability	16. Synergy between the dimensions	S-43	Creation of synergies by projects	The criterion is met when projects generate results in various dimensions of sustainability that combine to produce synergies.
		S-44	Identification of synergies by the evaluation	The criterion is met when the evaluation identifies potential for synergies.
	17. Conflict between the dimensions	S-45	Identification of conflicting objectives by the project	The criterion is met when conflicting objectives between dimensions are identified by the project.
		S-46	Identification of conflicting objectives by the evaluation	The criterion is met when the evaluation identifies conflicting objectives between dimensions.
	18. Side effects tolerable	S-47	Classification of possible compensation measures by the project as sufficient and / or of possible side-effects as 'tolerable'	The criterion is met when the project determines that compensation measures implemented (in order to minimise conflicting objectives between dimensions) are sufficient or that any side-effects generated by the project are 'tolerable'.
		S-48	Classification of possible side effects by the evaluation as 'tolerable'	The criterion is met when the evaluation determines that compensation measures implemented by the project are sufficient or that any side-effects generated by the project are 'tolerable'.

Source: Authors' own table

7.3
Team members

Core team	
Dr. Sven Harten	Head of Department
Dr. Martin Noltze	Senior Evaluator and Team Leader
Dr. Michael Euler	Evaluator
Ida Verspohl	Evaluator
Cornelia Michels-Lampo	Project Administrator

Team members	Position
Prof. Dr. Sebastian Vollmer	External peer reviewer
Dr. Kerstin Guffler	Internal peer reviewer at DEval
Solveig Gleser	Internal peer reviewer at DEval
Thomas Wencker	Internal peer reviewer at DEval
Jana Preiß	Associate master student
Niklas Witzig	Intern
Grisel Orozco	Intern
Helena Heberer	Student assistant
Sarah Stahlmann	Student assistant
Lea Smidt	Student assistant

7.4 Timeline

Concept phase		**Preparatory phase and definition of the object of the evaluation**
	04/2016 – 05/2016	Preliminary meetings with the BMZ and the implementing organisations
	06/2016 – 07/2016	Concept paper drafted
	08/2016	Meeting of reference group to discuss draft evaluation concept
	08/2016	Finalisation of the concept paper
Inception phase		**Development of the methodology**
	08/2016 – 10/2016	Inception report drafted
	10/2016	Meeting of the reference group to discuss the draft inception report
	02/2017	Finalisation of the inception report
Data collection and synthesis phase		**Data collection and analysis**
	10/2016 – 11/2016	Data and documents obtained from the implementing organisations
	11/2016	Establishment of dataset and sampling
	12/2016 – 02/2017	Procurement of secondary data as part of the evaluation synthesis
	12/2016 – 04/2017	Conduct of the quantitative content analysis
	02/2017	Conduct the contextual study and portfolio analysis
	03/2017 – 04/2017	Analysis and integration of the findings from the meta-evaluation and the evaluation synthesis
	05/2017	Meeting of the reference group for preliminary findings and conclusions
Reporting		**Production of the evaluation reports and dissemination**
	06/2017 – 07/2017	Drafting of the meta-evaluation and evaluation synthesis reports
	08/2017	Evaluation report forwarded to the reference group
	09/2017	Reference group meeting for presentation of the evaluation reports
	01/2018	Publication of the evaluation reports
	2018	Dissemination

German Institute for
Development Evaluation (DEval)

Fritz-Schäffer-Straße 26
53113 Bonn, Germany

Phone: +49 228 24 99 29-0
Fax: +49 228 24 99 29-904

E-mail: info@DEval.org
www.DEval.org

DEval
GERMAN
INSTITUTE FOR
DEVELOPMENT
EVALUATION