SUSTAINABILITY IN GERMAN DEVELOPMENT COOPERATION

Evaluation synthesis
2018

DEval
GERMAN INSTITUTE FOR DEVELOPMENT EVALUATION

This evaluation synthesis ‚Sustainability in German development cooperation' is part of DEval's thematic focus on sustainability. The evaluation synthesis is supported by an accompanying meta-evaluation. 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

Evaluation synthesis
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), FO Travel/Alamy Stock Foto (Chap. 1), themacx/iStock.com (Chap. 2), Nikon'as/Fotolia.com (Chap. 3), epicurean/iStock.com (Chap. 4 + 6), andresr/iStock.com (Chap. 5), Steve Bloom Images/Alamy Stock Foto (Chap. 7)

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

Printing
Bonifatius,
Paderborn

FSC www.fsc.org
MIX
Paper from responsible sources
FSC® C011558

ISBN 978-3-96126-065-2 (print)
ISBN 978-3-96126-066-9 (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 evaluation synthesis and the accompanying meta-evaluation. 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 proof-reader 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 undergraduate staff 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 makes sustainability the guiding principle for global action by humankind. The Sustainable Development Goals (SDGs) defined in the 2030 Agenda combine economic progress with social justice and the sound management of environmental resources. Responsibility for implementing the 2030 Agenda rests with all countries. At the same time, implementation requires new arrangements for cooperation between governments, the private sector, the scientific and academic community, and civil society.

The international development cooperation community has also pledged to reorient its approach accordingly. In the future, the design and implementation of development cooperation must comply with the goals and principles of the 2030 Agenda. This is a key challenge for international development cooperation. At the level of individual projects, it requires planners to reflect in particular on social, economic and environmental interactions, and effects on disadvantaged groups. To support this process, evidence-based recommendations are required. Currently there are only a limited number of projects that were designed explicitly in line with the 2030 Agenda and its principles. Nonetheless it is possible to study the sustainability of development cooperation projects empirically.

In evaluations of German development cooperation projects, sustainability has been systematically assessed since 2006. In that year the Federal Ministry for Economic Cooperation and Development (BMZ) published its 'Evaluation criteria for German bilateral development cooperation. A guideline for evaluations performed by the BMZ and the implementing organisations'. Based on the Principles for Evaluation of Development Assistance adopted by the Development Assistance Committee (DAC) of the Organisation for Economic Cooperation and Development (OECD) in 1991, this guideline contains instructions on assessing the evaluation criteria relevance, effectiveness, efficiency, impact and sustainability. Pursuant to the guideline, the sustainability of specific projects is assessed using mandatory key questions. The outcome of this assessment is the award of a sustainability score. Conceptually, the sustainability of projects is assessed in close conjunction with impact. It is therefore to be expected that evaluation practice to date – through the criterion of impact – already covers several of the principles of the 2030 Agenda.

The evaluation synthesis conducted here aims to better understand the interactions between various determinants when assessing the sustainability of projects. The purpose of the study is to help better align the strategic and operational orientation of German development cooperation with the new requirements of the modern understanding of sustainability contained in the 2030 Agenda. This is in response to the increased importance of sustainability when evaluating German development cooperation projects in conformity with the SDGs.

The present evaluation synthesis contains a first comprehensive and systematic aggregate assessment of sustainability in evaluations of German Financial and Technical Cooperation (FC and TC). The study is confined to evaluations of the two major official implementing organisations – the KfW Development Bank (KfW) and the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH. 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. In addition to purely bilateral projects in specific countries, the study also covers regional, sectoral and global projects.

Methodology

The factors affecting the sustainability score were analysed using multivariate regression models. These models allow investigators to ascertain the effect of various factors on the variable to be explained – in this case the score awarded for the sustainability of projects. Due to the limited availability of data it was only possible to include certain factors. Consequently the study is restricted to specific features of projects, factors associated with their implementation and available contextual information. The latter include both specific features of the immediate context of the development projects, and macro quantitative indicators at the level of partner countries. Furthermore, the analysis also draws on findings of the accompanying meta-evaluation on sustainability in German development cooperation. The findings of the meta-evaluation allowed the evaluation team to include in

their analyses the criteria used to assess sustainability. Secondly, the investigators used the assessment of evaluation quality performed in the meta-evaluation as a weighting factor for individual observations in the regression models. No observations were ruled out of the analysis. However, the weighting of individual observations does ensure that the most credible findings received the greatest weighting in the synthesis.

Key findings, conclusions and recommendations

Factors affecting the rating of project sustainability

In the evaluations conducted by the KfW and GIZ the sustainability score varies only slightly. Over 84 per cent of the evaluations included awarded a score of 2 or 3 for sustainability. Furthermore, the higher the score awarded for the DAC criteria relevance, effectiveness, efficiency and impact, the higher the sustainability score (1 = highest score, 6 = lowest score). Consequently, in all regression models the average score for all DAC criteria (excluding sustainability) – according to statistical level of significance and effect size – is the key determinant of the sustainability score.

Hence sustainability is an overarching evaluation criterion. It contains barely any genuine determinants that can be strictly separated from the remaining DAC criteria. Nonetheless, the regression models do demonstrate that certain factors are particularly important with regard to sustainability rating. In particular, the information obtained from the accompanying meta-evaluation on sustainability permits conclusions regarding the sustainability of specific projects. The findings of the accompanying meta-evaluation also demonstrate that although sustainability is assessed on the basis of comprehensive criteria in practice, this assessment is at the same time performed unsystematically and inconsistently. Through the assessment of impact, the assessment of sustainability is also always linked to the assessment of the other DAC criteria.

Differences in the assessment of sustainability also arise according to the type of evaluation used. While ex-post evaluations base their assessments on observations, in project progress reviews (PPRs), project evaluations (PEs) and final evaluations sustainability is assessed on the basis of a prognosis. Compared to the other types of evaluation, ex-post evaluations tend to award the lowest scores for project sustainability. But it is not only the scores that differ depending on the type of evaluation. So too do the criteria on which they are based. Comparing the sustainability scores between different projects is thus only possible to a limited extent. Generally speaking, however, it can be concluded that in ex-post evaluations the role and the contributions of development partners and target groups are particularly important for the sustainability of projects. By contrast, when sustainability is assessed in PPRs, PEs and final evaluations it is primarily the direct outputs, the implementation of the project and the context of implementation that are taken into account.

Alongside these differences, however, the determinants identified in the different types of evaluation also display commonalities. For instance, in both ex-post evaluations and in PPRs, PEs and final evaluations, the predictability of the continuation of results has a significant positive effect on project sustainability. This shows that in all types of evaluation, not only the outputs and results of projects, but also the durability of results – a key conceptual element in the assessment of sustainability – has a significant effect on the sustainability score.

Recommendations on boosting the sustainability of projects

The recommendations below result from the findings and conclusions of the evaluation synthesis. Due to their complexity, the recommendations are supplemented – in the various sub-points – by suggestions and ideas that relate primarily to their application.

The evaluation team recommends that when planning and implementing projects, the BMZ and the implementing organisations should take greater account of the capacities of the partners and executing agencies on the ground, and systematically support their development.

- With this in mind, an explicit assessment of the capacities of all relevant partners and agencies might also be taken into consideration when deciding on the eligibility for support of a module during project

planning. Here it should be ensured that the partners and agencies possess the technical, financial and institutional capacities to continue the activities and outputs previously generated by the project.
- Furthermore, the capacities of the partners and agencies could be analysed repeatedly at regular intervals in the course of an ongoing project. Successfully transferring the outputs to the partners at the end of the project could also be underpinned by developing long-term exit strategies.
- Strengthening the partner system might ensure partner-country ownership of implementation of the 2030 Agenda.

The evaluation team recommends that the GIZ and KfW in future understand the factors relevant to the management of the project not only in relation to effectiveness, but also in direct relation to sustainability, and take this into account accordingly.

- These include particularly the use of institutional structures on the ground, the systematic analysis of lessons learned and the development of scaling-up and exit strategies.

Systematic learning from evaluations

The comparability of evaluation findings is a key prerequisite for conducting evaluation syntheses. Aggregating findings from individual evaluation reports promotes systematic, strategic and cross-institutional learning. Unfortunately, the findings on the sustainability of development cooperation projects found in the evaluation reports are only comparable to a certain extent. There are various reasons for this.

First of all, although the key questions do provide guidance for assessing sustainability, they are not sufficiently operationalised. This is reflected by the fact that the specific criteria underlying each individual score are manifold, and cannot always be specified unequivocally. Given the diversity of the portfolio of implemented measures a certain flexibility in assessment is necessary; even so, the assessment of sustainability must also be comprehensible and comparable for outsiders. This idea is also reflected in the principle of joint accountability in the 2030 Agenda.

Secondly, the implementing organisations studied here display systematic differences in the practice and management of evaluation. The findings demonstrate that GIZ evaluations award significantly higher sustainability scores than KfW evaluations – even though the same number of criteria are rated positively. Furthermore, the use of different types of evaluation both within and between the implementing organisations leads to structural differences in the assessment of sustainability. There are also fundamental differences in the way the two implementing organisations manage evaluations. At the KfW all ex-post evaluations are audited by the evaluation department. Here the assessment of individual measures is placed in the context of the assessment of comparable measures. By contrast, the conduct of PPRs and PEs is decentralised. Responsibility rests with the officer responsible for the commission in question. Whereas at the KfW a core team of staff members checks all reports, thus establishing a minimum degree of comparability, the decentralised evaluation system at the GIZ precludes the organisation-wide comparison of individual reports. It is therefore to be assumed that overall, evaluations of GIZ projects are more heterogeneous and depend more heavily on attributes of the authors than is the case at the KfW.

Thirdly, the meta-data from evaluations and projects that are recorded by the implementing organisations tally only to a certain extent. Information relevant to the present analysis was in some cases incomplete, or was systematically recorded by only one implementing organisation.

The sketchy comparability of sustainability ratings makes it more difficult to identify enabling factors for sustainability. For instance, on the basis of the information available it is not possible to establish definitively whether the macroeconomic and political indicators integrated into the models actually have no effect on the sustainability of projects, or whether it is not possible to establish any link at all due to the lack of comparability and transparency of the criteria on which the assessment was reached. The potential for obtaining from evaluation syntheses strategic findings and findings that

would be relevant to the management response is thus very limited.

Recommendations on boosting systematic, strategic and cross-institutional learning
The recommendations below are also supplemented with suggestions and ideas that relate chiefly to their application.

To guarantee the systematic assessment of sustainability, the evaluation team recommends that the BMZ and the implementing organisations develop standardised and binding criteria. These should serve as a basis for the award of scores, and should be weighted transparently for this purpose.

- To take due account of the heterogeneous portfolio of German Technical and Financial Cooperation, the criteria should possess an appropriate degree of sector- and region-specific flexibility. Binding instructions on applying the criteria might also be defined separately for each sector or for TC/FC modules.

The evaluation team recommends that the BMZ and the implementing organisations – where possible – harmonise meta-data on projects and their evaluations and record this information at a central point.

- The systematic and central recording of meta-data from projects and evaluations would make cross-institutional, aggregated analyses considerably easier to perform, and therefore quicker.
- With this in mind, the BMZ and the implementing organisations might explore how they could meet the requirements of joint accountability articulated in the 2030 Agenda by recording and systematically preparing meta-data.

CONTENTS

Figures

Tables

ABBREVIATIONS AND ACRONYMS

BMZ
German Federal Ministry for Economic Cooperation and Development

DAC
Development Assistance Committee of the OECD

FC
Financial Cooperation

GDP
Gross domestic product

GIZ
Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

KfW
KfW Development Bank

ODA
Official Development Assistance

OECD
Organisation for Economic Co-operation and Development

PE
Project evaluation

PPR
Project progress review

SDGs
Sustainable Development Goals

TC
Technical Cooperation

1.

INTRODUCTION

This evaluation synthesis represents a first comprehensive empirical study of the sustainability of German bilateral development cooperation projects and the factors affecting it. It is based on evaluations performed by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) and KfW Development Bank (KfW) on projects financed through public development funds of the German Federal Ministry for Economic Cooperation and Development (BMZ).

1.1 Background

The success of development cooperation is measured by the sustainability of its results. The launch of the 2030 Agenda for Sustainable Development made sustainability the guiding principle for action by humankind. All countries are responsible for implementing the 2030 Agenda. International development cooperation must also refocus its approach. At the overarching level the key issues are the coherence of development cooperation with other policy fields, the establishment of partnerships between governments, the private sector, civil society, and the scientific and academic community, and the mobilisation of funds to achieve the Sustainable Development Goals (SDGs) defined in the Agenda. At the level of individual development cooperation projects, the 2030 Agenda affects their design, planning and implementation. Here the key issue is how to ensure the sustainability of the results generated by individual projects as envisaged in the 2030 Agenda. The Agenda envisions projects that take into account interactions between the social, economic and environmental dimensions, and include disadvantaged groups. Planning and implementing projects in conformity with this vision is a key challenge for international development cooperation. To support this process, evidence-based recommendations are required. To the best of the evaluation team's knowledge, currently there are only a limited number of projects that were designed explicitly in line with the 2030 Agenda and its principles. Nonetheless it is possible to study the sustainability of development cooperation projects empirically.

In evaluations of German development cooperation projects, sustainability has been systematically assessed and scored since 2006. In that year the Federal Ministry for Economic Cooperation and Development (BMZ) published its 'Evaluation criteria for German bilateral development cooperation', which contains mandatory key questions for assessing the sustainability of individual development cooperation projects (BMZ, 2006).[1] According to the guideline, sustainability is to be assessed on the basis of the continuation of development results, stability of the context in terms of social justice, economic performance, political stability and ecological balance, as well as the risks and potentials for (lasting) effectiveness (BMZ, 2006). Conceptually, the sustainability of projects is assessed in close conjunction with impact. A meta-evaluation accompanying this evaluation synthesis demonstrates that in practice, the assessment of sustainability actually involves several evaluation criteria, and that sustainability is therefore being understood in a comprehensive sense, and evaluated and assessed accordingly (Noltze et al., 2018). It is therefore to be expected that the existing practice of evaluation already covers several of the principles of sustainable development as envisioned in the 2030 Agenda. A systematic analysis of the factors affecting the sustainability score therefore offers an opportunity to obtain relevant findings for the design of development cooperation projects in the age of the 2030 Agenda.

1.2 Purpose of the evaluation synthesis

The purpose of this evaluation synthesis is to comprehensively and systematically analyse how the sustainability of German development cooperation projects is being assessed. By identifying key factors influencing the sustainability score, the study aims to bring to light possible ways of making German development cooperation projects more sustainable as envisaged by the 2030 Agenda. Using statistical models, the study investigates the extent to which factors at the level of the evaluation reports, the projects evaluated, and the country in which the project was implemented, affect the sustainability score.

Apart from using meta-data from projects and evaluation reports – such as project duration and volume of funding – the analysis draws on the findings of the accompanying meta-evaluation on sustainability (see Noltze et al., 2018). Based on

1 *It also includes binding instructions on assessing the evaluation criteria relevance, effectiveness, efficiency and impact.*

evaluation reports prepared by the GIZ and KfW, the meta-evaluation identifies the criteria that were used to assess sustainability.[2] This evaluation synthesis provides an opportunity to better understand the interactions between various determinants of the sustainability score. In the context of the 2030 Agenda, it can therefore help align German development cooperation more closely with multidimensional sustainability, both strategically and operationally. This is in response to the increased importance of sustainability when evaluating projects in German development cooperation in conformity with the SDGs. It is assumed that although the assessment of sustainability by evaluations is not an objective measure of the sustainability of German development cooperation projects, it is the best possible approximation of such. The motivation for this analysis is the introduction of the 2030 Agenda for Sustainable Development, and its emphasis on sustainability as the key element of the debate on effectiveness.

1.3 Object

The object of the evaluation synthesis is the sustainability of German development cooperation projects and the factors affecting it. Specifically, the analysis focuses on the aggregate assessment of sustainability in evaluations of German Financial and Technical Cooperation. In evaluations, the sustainability of a project is expressed as a score, the determinants of which are subjected to statistical analysis here. As the object of the evaluation synthesis was to be addressed as comprehensively as possible, the study is not restricted either to particular sectors, or to particular regions or types of project. In addition to purely bilateral projects in specific countries, the study also covers regional, sectoral and global projects.

This first systematic analysis of the sustainability of projects is restricted, however, to evaluations by the two major official implementing organisations – the KfW and GIZ.[3] Every year the two implementing organisations deliver a significant portion of public development finance, and have a sectoral and regional portfolio that is highly diversified. At the same time both implementing organisations have a high degree of evaluation coverage of individual projects (today referred to as modules). Since 2006, sustainability has been assessed in all evaluations as a criterion of performance by German development cooperation. The assessment is based on the BMZ guideline on how to apply the DAC criteria. This evaluation synthesis therefore includes only evaluations that were conducted and completed between July 2006 and the point at which the data were collected in October 2017.

When ascertaining the determinants it becomes necessary to restrict the object of the analysis due to the limited availability of data. The analysis is restricted to specific features of projects, factors associated with their implementation and available contextual information. The contextual factors include both specific features of the immediate context of the development projects, and macro quantitative indicators at the level of partner countries. An increase in the availability of data and the broadening of the object of the evaluation was facilitated by the accompanying meta-evaluation on the practice of evaluating the sustainability of projects (Noltze et al., 2018).

[2] *The methodology and findings are described in Noltze et al. (2018).*
[3] *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.*

1.4 Evaluation questions

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

Evaluation question 1 – What specific features become evident when taking an overall look at sustainability in the portfolio of evaluations in German development cooperation?

Evaluation question 2 – To what extent do project-specific factors affect the sustainability score of development projects?

Evaluation question 3 – To what extent do context-specific factors affect the sustainability score of development projects?

Evaluation question 4 – To what extent do the underlying assessment criteria affect the sustainability score of development projects?

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

1.5 Structure of the evaluation report

The evaluation synthesis is structured as follows:

Chapter 2 presents the practice of evaluating and assessing sustainability in German development cooperation projects (Section 2.1 and Section 2.3). There, possible factors affecting sustainability are identified, and their theoretical link to the sustainability of projects is discussed (Section 2.2). The section concludes by describing the database (Section 2.4) and the sampling procedure for the present analysis (Section 2.5).

Chapter 3 describes the methodology of the evaluation. In addition to the empirical strategy (Section 3.1), different forms of statistical modelling are discussed (Section 3.2) and limitations and challenges are identified (Section 3.3).

The findings of the evaluation synthesis are presented in Chapter 4. The section on findings begins by describing the explanatory variables (Section 4.1), and discusses the sustainability score as the dependent variable of the analysis (Section 4.2). Finally, the findings are presented in relation to the evaluation questions (Section 4.3).

The conclusions and recommendations are contained in Chapter 5.

2.

SUSTAINABILITY IN GERMAN DEVELOPMENT COOPERATION

This Chapter will first of all discuss which key questions are used to assess the criterion of sustainability. In this connection the report points to the limitations of assessing sustainability. It also discusses possible factors affecting the sustainability score awarded. It then goes on to discuss the practice of evaluation by the GIZ and KfW. Finally the report presents the database of this evaluation synthesis and the distribution of sustainability scores awarded across the portfolio of reports.

2.1 Assessing sustainability in German development cooperation

Since 2006, the sustainability of German development cooperation projects has been systematically assessed in all evaluations performed by the BMZ and its implementing organisations. The latter are carried out on the basis of a guideline for these evaluations (BMZ, 2006). Based on the DAC Principles for Evaluation of Development Assistance (OECD, 1991), the guideline contains instructions on assessing the DAC criteria relevance, effectiveness, efficiency, impact and sustainability.

According to the guideline, sustainability is to be assessed in relation to three key aspects. The first is the continuation of development results over time. The second is the stability of the project context with respect to the factors of social justice, economic performance, political stability and ecological balance. Thirdly, sustainability is to be assessed in relation to the risks and potentials for the project's continued effectiveness (BMZ, 2006).[4]

The outcome of this assessment is the award of a score of between 1 and 4 (1 = highest score, 4 = lowest score).[5] 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 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. A product is considered 'sustainable' when it is awarded a score of between 1 and 3. Projects awarded a score of 4 are considered 'unsustainable'.[6]

On closer inspection, two things are striking about the individual scores. First of all, although projects with a score of 3 are formally rated as 'sustainable', a score of 3 also means that the project's impact is either inadequate or is expected to diminish significantly. Strictly speaking, this definition would mean that projects with a score of 3 could just as well be rated as 'unsustainable'. Secondly, the definitions for all the scores clearly indicate that there is a conceptual link between a project's impact and its sustainability. If a development project has no positive impacts, it cannot be sustainable. So far, however, this link has remained implicit in the conceptual rationale of the DAC criteria. On its own it does not yet give clear guidance on how to deal with sustainability in evaluations. With this in mind, the accompanying meta-evaluation analyses empirically how sustainability is actually understood, evaluated and assessed in practice (Noltze et al., 2018). Here it emerged that in evaluations, project sustainability is indeed being examined, discussed and assessed on a conceptually comprehensive and complex basis, albeit at the same time unsystematically and inconsistently. This finding demonstrates that the sustainability score awarded in evaluations contains much more information than the key questions contained in the BMZ guideline would initially lead one to assume. For the present evaluation synthesis this finding is extremely important, because analysing the multidimensional concept of sustainability ultimately requires evaluators to take manifold factors into account, and therefore entails a high overall data requirement. This is why the evaluation synthesis also includes in its analysis additional information from the accompanying meta-evaluation.

4 *These key questions go beyond the OECD-DAC's understanding of sustainability in that sustainability is mainly defined as the continuation of development results once a project has come to an end. The full definition of the OECD-DAC criteria can be found here: http://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm .*

5 *In GIZ's so-called project evaluations (PEs), which were introduced in April 2014, sustainability is assessed along a six-point scale.*

6 *In the overall assessment, a project is only considered to be successful, if its sustainability is rated as being at least satisfactory (score of 3). This also applies to the criteria 'effectiveness' and 'impact'.*

2.2
Factors affecting the sustainability score

What makes development cooperation projects sustainable? The existing literature on sustainability in development cooperation answers this question only to a limited extent. It focuses above all on an overarching conceptual discussion, rather than the sustainability of individual projects. This means it deals primarily with the importance of sustainability for development cooperation and the challenges of sustainable development. The articles contained in the anthology by König and Thema (2011) entitled Nachhaltigkeit in der Entwicklungszusammenarbeit [Sustainability in development cooperation], for instance, highlight the importance of sustainability for development cooperation. They also critically discuss the concepts of 'sustainability' and 'sustainable development', emphasise the role of the global financial and trade order for sustainable development, discuss issues concerning the coherence of development policy, and shed light on lessons learned in the evaluation of sustainability in German Financial Cooperation projects. Caspari (2004) highlights the complexity of the evaluation criterion 'sustainability', and develops a conceptual framework for assessing it consistently. Contributors to the anthology edited by Raggamby and Rubik (2012) discuss inter alia how the evaluation of sustainability can contribute towards policy formulation. They also describe indicators that are relevant to policy as well as methods for evaluating sustainability. And they highlight quality standards that evaluations should meet when assessing sustainability. A number of more recent publications focus on the importance of evaluation for implementing the 2030 Agenda. Some authors suggest that national policies for achieving the SDGs should be monitored through national evaluation systems (Benoit et al., 2017; Ofir et al., 2016). Another proposes that the evaluation agenda of specific countries should pursue a holistic approach, and assess policies and projects not in isolation, but in the wider national context (Ofir et al., 2016). Other authors argue that in order to take due account of the complexity of the 2030 Agenda, we should go beyond merely monitoring indicators. In particular, they suggest that rigorous impact evaluations should identify why, how and under what conditions policies generate results, and which groups benefit from them (Lucks et al., 2016; Schwandt et al., 2016). Furthermore, one study argues that a broad range of stakeholders should be included in order to create a country-specific focus on individual indicators (Lucks et al., 2016).

To the best of the evaluation team's knowledge, to date there are no empirical findings on the factors that influence assessment of the sustainability of specific projects. However, there are a number of studies that analyse the factors affecting assessment of the overall performance of projects of the World Bank, as well as the African and Asian development banks. A synopsis of the studies reveals that the performance of a project is influenced primarily by the specific characteristics of the project and its modes of implementation, the characteristics of its evaluation, and contextual factors at the country level (Assefa et al., 2014; Bulman et al., 2015; Denizer et al., 2013; Dollar and Levin, 2005; Kilby, 2013). Since the overall performance of a project is crucially dependent on its sustainability, the present study explores the extent to which the factors identified also affect the assessment of sustainability.

Regarding the factors at country level, Denizer et al. (2013), show that the performance of a project is positively affected by a country's economic development status and its economic stability. The evaluation synthesis therefore examines whether economic development also has positive effects on the sustainability of a project. Positive economic development may for instance lead to an increase in public revenues, which in turn increases the scope for the partner country to contribute financial or human resources for the implementation of projects (Bulman et al., 2015; Denizer et al., 2013; Hemmer and Lorenz, 2003). The political rights and civil liberties of a society also correlate positively with the performance of projects (Bulman et al., 2015; Denizer et al., 2013; Dollar and Levin, 2005; Isham et al., 1995). Furthermore, a higher level of rule of law and democracy within a country is also conducive to the performance of projects implemented there (Chauvet et al., 2010; Denizer et al., 2013; Dollar and Levin, 2005). The rule of law encourages investment, because it creates a higher degree of trust among different stakeholders and lowers transaction costs (Dollar and Levin, 2005). Democratic institutions that work encourage governments to be publicly accountable. Governments then face pressure from their electorate, which

strengthens their interest in implementing effective projects (Bulman et al., 2015; Denizer et al., 2013; Dollar and Levin, 2005; Isham et al., 1995). It seems plausible that the rule of law and level of democracy will also have a positive effect on the sustainability of projects.

Several studies have found that compared to factors at the country level, factors at the level of projects have a relatively strong influence on overall project performance (Bulman et al., 2015; Denizer et al., 2013; Dollar and Levin, 2005; Isham et al., 1995). According to these studies, factors affecting project performance are the amount of funding, project duration and the sector involved. Here we should note that longer and more costly projects are not necessarily rated more favourably (Bulman et al., 2015; Denizer et al., 2013; Dollar and Levin, 2005; Isham et al., 1997). There may be a possible link between a project's duration and volume of funding, and its complexity. The overall rating might then be a product chiefly of the complexity of the system of objectives (Bulman et al., 2015; Denizer et al., 2013; Dollar and Levin, 2005; Isham et al., 1995). A longer period for preparing the implementation of a project (Dollar and Levin, 2005; Kilby, 2013), and a higher level of managerial expertise, make it more likely that the project will perform well (Chauvet et al., 2010; Denizer et al., 2013; Dollar and Levin, 2005). By contrast, a delay in project implementation may have a negative effect on project performance (Chauvet et al., 2010; Denizer et al., 2013; Dollar and Levin, 2005).

The rating of project performance is also determined by specific features of the evaluation. Denizer et al. (2013) show that the score awarded becomes poorer, the longer the interval between the end of the project and the date of the evaluation. A similar link in relation to the sustainability of projects is also plausible. The later the project impacts are evaluated after the project has come to an end, the more likely it is that these impacts will have diminished. Moreover, there is a possible link between rating practice and the quality of the methods used in the report. It cannot be ruled out that methodologically superior (or inferior) evaluations score the sustainability of projects more discerningly (or less discerningly), and therefore award a lower (or higher) score.

All of the above-mentioned factors form exclusively the influence exerted by the implementation context and the descriptive characteristics of a project and its evaluation. The assessment of sustainability is also determined by the results of the project and its implementation, however. In the studies quoted here, these aspects have only been dealt with to a limited extent. This is presumably due to the poor availability of relevant information. Data on the achievements of specific development cooperation projects that are linked to the assessment of their sustainability can only be obtained directly from the project documents. To close this gap, this evaluation synthesis draws on the findings of the accompanying meta-evaluation (Noltze et al., 2018). The latter developed a conceptual analytical framework for recording the criteria used when assessing sustainability. According to this framework, the assessment of sustainability is determined by seven areas: the context of the measure, its implementation, the results/outcome achieved, the local capacities, the unintended effects (impact) [7] of the project, the predictability of the continuation of results over time and the interaction between the dimensions.[8]

[7] *'Impact' includes both 'intended' and 'unintended' effects. However, since the 'intended effects' are an integral part of the assessment of the OECD-DAC impact criterion, this study will look only at the 'unintended' effects, which play a special role conceptually in the assessment of sustainability. Noltze et al. study the intended effects (2018).*

[8] *For a detailed discussion of the analytical framework for the assessment of sustainability, see Noltze et al. (2018).*

2.3 Evaluation practice of GIZ and KfW

Since 2006, the key criteria for evaluations conducted by the GIZ and KfW have been prescribed on a mandatory basis in a guideline issued by the BMZ (BMZ, 2006). Selection of the specific reporting format and the conduct of evaluations are the responsibility of the respective implementing organisations. When assessing specific projects, the GIZ and KfW use different types of evaluation.

Since 2006, the GIZ has been using both centralised and decentralised evaluations to evaluate specific projects. The centralised evaluations were managed, and up to and including 2014 used, by the GIZ's Evaluation Unit. These were conducted independently of the implementation of the projects under evaluation (i.e., they were independent evaluations). The independent evaluations included ex-ante, interim, final and ex-post evaluations. The ex-ante and interim evaluations were conducted prior to or during the course of the project, whereas the final evaluations were usually conducted six months before the end or after the end of the project, and ex-post evaluations were conducted two to five years after completion of the project. Independent evaluations were conducted for projects in a specific sector on an annually rotating basis. The decentralised evaluations included so-called project progress reviews (PPRs), which were used until March 2014. Since April 2014 these have been replaced by so-called project evaluations (PEs). Project evaluations are now the only remaining type of evaluation for assessing individual projects. Unlike the centralised types of evaluation, responsibility for implementing the decentralised evaluations rests with the respective officer responsible for the relevant project commission. PPRs and PEs are carried out six to twelve months before the end of projects.[9]

Unlike the GIZ, the KfW has been evaluating individual projects using ex-post evaluations throughout since 2006. KfW's ex-post evaluations are usually implemented three to five years after the end of a project. At the KfW, evaluations are organised by the independent evaluation unit of the KfW Development Bank. Since 2006, the selection of projects for evaluation has been based on an annual random sample of completed projects that includes half the projects within each sector.10

With regard to their assessment of sustainability, the individual types of evaluation are only comparable to a limited extent. PPRs, PEs and final evaluations are conducted immediately upon completion of a project. De facto, assessing the sustainability of project results achieved involves assessing future developments. By contrast, in ex-post evaluations the assessment of sustainability is based on observations and actual developments that extend at least three years beyond the end of the project. These differences need to be taken into account when analysing the factors that influence the sustainability scores of projects.

2.4 Database and portfolio analysis

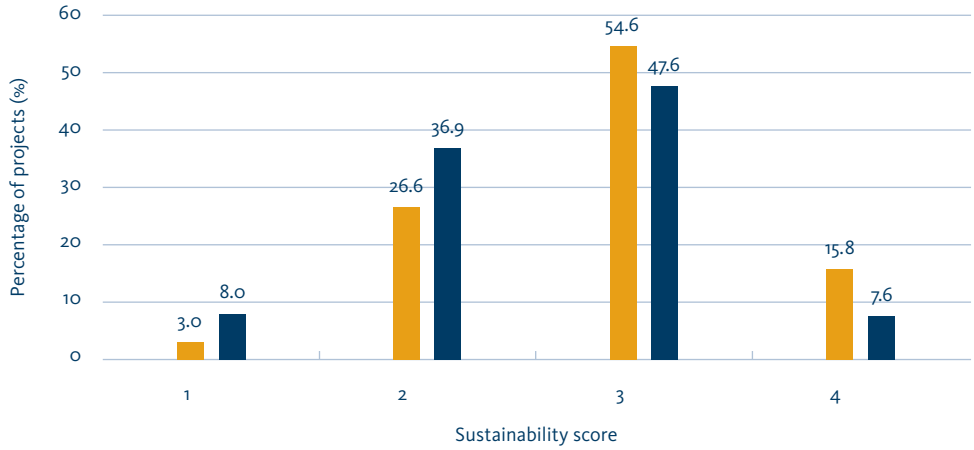
The database (observations) for the present report comprises GIZ and KfW projects that were evaluated between 2006 and 2016 using the DAC criteria.[11] When the data were collected in October 2016, a total of 1,015 evaluated projects were included in the population.[12] Of these, 462 involved Financial Cooperation (KfW) and 553 Technical Cooperation (GIZ). While all the KfW evaluations are ex-post evaluations, the GIZ evaluations break down into 56 ex-post, 44 final and 343 project evaluations, plus 110 project progress reviews. As well as bilateral projects, the population also includes so-called sector, regional and global projects.[13]

Figure 1 shows the sustainability score awarded by implementing organisation for all evaluations included in the population.[14]

Regarding interpretation of the graphic, the reader is referred to the description of the scores in section 2.1. As described

9 *For a detailed description of the GIZ's evaluation system, please visit: https://www.giz.de/en/aboutgiz/monitoring_and_evaluation.html*
10 *For a detailed description of the KfW's evaluation system, please visit: https://www.kfw-entwicklungsbank.de/International-financing/KfW-Development-Bank/Evaluations/*
11 *Since ex-ante and interim evaluations are conducted relatively early during the life of a project, they would appear to be unsuitable for assessing sustainability understood as meaning the durability and stability of results. Both types of evaluation were therefore excluded from the population.*
12 *We should note that GIZ and KfW often comprise a chronological sequence of phases involving continuity of content (referred to as 'modules'). While final and ex-post evaluations are not followed by a further phase (or module) of the project, when a project progress review or a project evaluation is carried out there may be a further phase or module of the project, and therefore a subsequent evaluation. To capture the latest possible assessment of sustainability, the population includes only the most recent evaluation of each project.*
13 *The population includes 99 regional projects (87 of the GIZ, 12 of the KfW), 52 sector projects (35 of the GIZ, 17 of the KfW) and 6 global projects (of the GIZ).*
14 *In addition to the figures shown below, Table 8 describes the characteristics of the population by implementing organisation.*

Figure 1: Sustainability score awarded by implementing organisation



Source: Authors' own graphic

Notes: The graphic shows the relative frequencies of the sustainability scores awarded by implementing organisation. It includes all the evaluated projects in the population. N = 1.015

there, a score of up to 3 attests to the fact that the positive development results of the project in question will either prevail for the foreseeable future, or demonstrably continue after the end of the project. This assessment is reached in 93 per cent of all GIZ projects and 85 per cent of all KfW projects. In other words, around nine out of ten development cooperation projects are classified by their evaluations as 'sustainable'.[15]

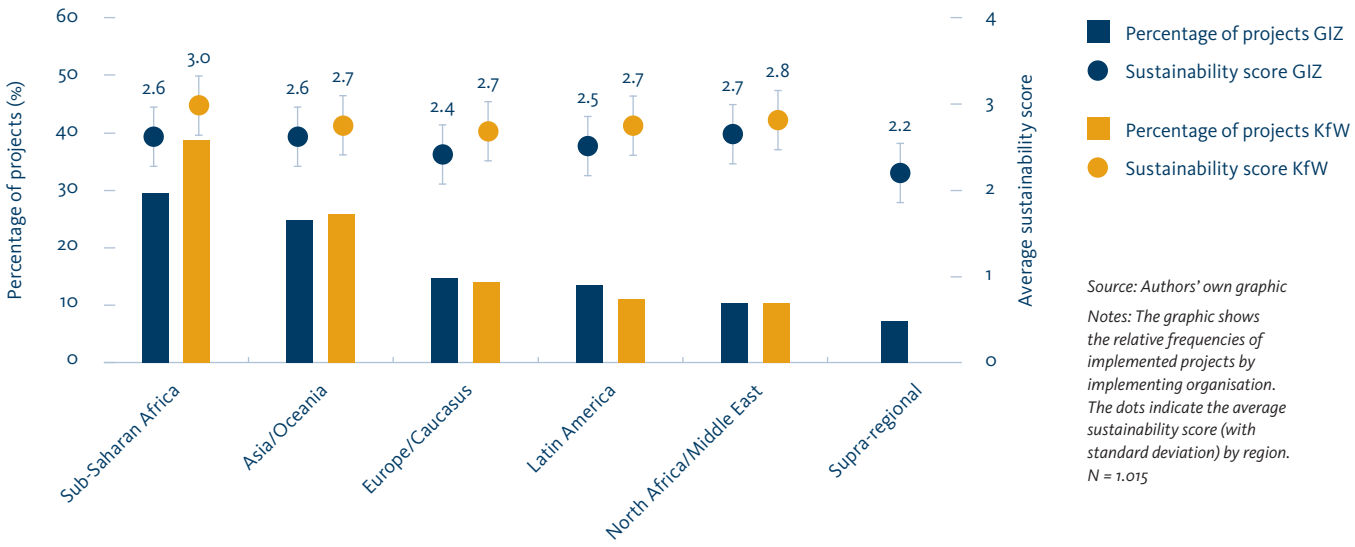
The GIZ and KfW portfolio referred to here includes projects from four continents and ten sectors. Figure 2 shows the distribution of these projects across various regions as well as the average sustainability score awarded by region and implementing organisation. The bars show the relative frequency of implemented projects, while the dots represent the average score awarded. As the graphic shows, both implementing organisations implement the majority of their projects in sub-Saharan Africa. The percentage of KfW projects in Africa is significantly higher than the corresponding figure for the GIZ.[16] Projects in the regions Asia/Oceania, Europe/Caucasus, Latin America and North Africa account for a similar percentage of the portfolio for both implementing organisations. GIZ implements a small percentage of its projects at the global (i.e. supra-regional) level (in sector and global projects).

Regarding project sustainability rating, it emerges that the average sustainability score awarded for KfW projects is lower across all regions compared to GIZ projects. Statistically significant differences between the scores for the two implementing organisations exist only in the sub-Saharan Africa and Europe/Caucasus regions, however.[17] Within the GIZ's portfolio supra-regional projects receive the best sustainability ratings. These projects receive significantly higher scores than GIZ projects in sub-Saharan Africa, Asia/Oceania and North Africa/Middle East.[18] Supra-regional projects differ from bilateral projects in that they cannot be assigned to a specific partner country. This means they are less dependent on implementation structures. Within the KfW portfolio projects in sub-Saharan Africa are rated significantly less favourably than projects in Europe/Caucasus and Asia/Oceania.[19]

Figure 3 shows the sectoral distribution of projects and the average sustainability score awarded by sector. Once again the data are presented separately for GIZ and KfW projects.

[15] *If – as described in Section 2.1 – we were to classify projects awarded the score 3 as 'unsustainable', only around 45 per cent of GIZ projects and around 30 per cent of KfW projects would be 'sustainable'.*
[16] *This statement is based on a two group proportion test.*
[17] *Due to the normal distribution of the score and homogeneous variance within the region, this statement is based on a variance analysis (ANOVA).*
[18] *This statement is based on a variance analysis (ANOVA).*
[19] *This statement is based on a variance analysis (ANOVA).*

Figure 2: Regional distribution of projects and their sustainability rating by implementing organisation



Source: Authors' own graphic

Notes: The graphic shows the relative frequencies of implemented projects by implementing organisation. The dots indicate the average sustainability score (with standard deviation) by region. N = 1.015

The findings demonstrate that the sectors sustainable economic development and government and civil society are particularly significant for both implementing organisations. Projects are also implemented relatively frequently in the sectors water and health. There are significant differences between the GIZ and the KfW with respect to the sector portfolio.[20] For example, the percentage of GIZ projects in the sustainable economic development, government and civil society, environment and education sectors is significantly higher compared to the KfW portfolio. By contrast, the KfW implements a significantly higher proportion of its projects in the water, health, energy, agriculture and transport sectors. These differences reflect the different core competences of the two implementing organisations. The GIZ performs Technical Cooperation (TC), for instance, and is usually actively involved in implementation in the partner country. The KfW on the other hand performs chiefly Financial Cooperation (FC), and focuses for the most part on (promoting) investment and dialogue with partners.

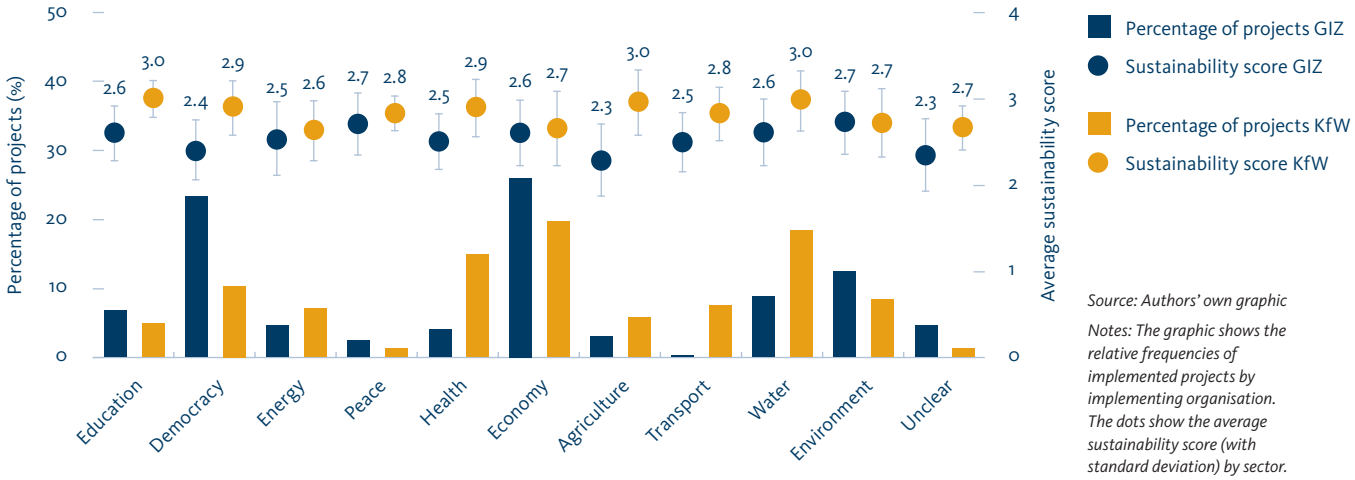
Figure 3 also shows that the sustainability score varies between sectors only moderately. Within the GIZ portfolio the government and civil society and agriculture sectors receive the best scores. In the case of the KfW, projects in the energy sector are rated as particularly 'sustainable'. With the GIZ projects in the peace and environment sectors receive the worst scores, while the KfW receives its worst scores for projects in the education, agriculture and water sectors. Within the portfolios of both implementing organisations there are no significant differences in scores between the individual sectors.

2.5 Sampling procedure

When analysing the factors affecting the sustainability score, we could conceivably take all observations in the population into account. A larger number of data points would allow us to determine links between the sustainability score awarded and individual factors with a higher degree of statistical certainty. However, for this population meta-data are available only for the project characteristics and the evaluation reports. We would not be able to use the meta-data available to determine the effects of the assessment criteria used in the reports and

[20] *This statement is based on a two group proportion test.*

Figure 3: Sectoral distribution of projects and their sustainability rating by implementing organisation



Source: Authors' own graphic

Notes: The graphic shows the relative frequencies of implemented projects by implementing organisation. The dots show the average sustainability score (with standard deviation) by sector. N = 1.015

the methodological quality of the reports on the sustainability score. Hence we would not be able to fully answer the evaluation questions mentioned at the outset using the meta-data alone.

We therefore draw on information from the accompanying meta-evaluation by Noltze et al. (2018). The meta-evaluation records the criteria used to assess sustainability in the various reports. This is done with the aid of an analysis grid comprised of seven areas. The individual areas are broken down into a total of 18 criteria and 48 differentiated criteria. The meta-evaluation also assesses the methodological quality of the reports. This too is performed using an analysis grid. The grids to record the criteria for assessing sustainability and for assessing methodological quality are included in the Annex (Table 6 and Table 7). The meta-evaluation was performed for a sample of the existing GIZ and KfW evaluation reports. Due to the differences in the assessment of project sustainability discussed in Section 2.3 , sampling was performed separately for each type of report.[21]

Table 1 shows the number of observations in the population per evaluation type, and the sample analysed by Noltze et al. (2018) . When determining the sample size the distribution of scores within the population of each evaluation type was taken into account. Based on the distribution of the sustainability scores, and the percentage of 'sustainable' projects (scores 1 to 3) and 'unsustainable' projects (score 4), two different sample sizes were first of all calculated. For each evaluation type the larger sample was included in the meta-evaluation (Noltze et al., 2018). The average sustainability score awarded, the percentage of projects rated 'sustainable' per evaluation type and the individual sample sizes are shown in Table 9 in the Annex.

The sample for the meta-evaluation also forms the basis for the empirical analyses performed here. The sample includes a total of 513 evaluated projects, of which 341 were GIZ projects and 172 KfW projects. Due to the relative frequencies of the evaluation types within the population, and the respective distributions of scores, the sample is made up of differing

[21] *In meta-evaluations conducted by more than one person the findings can be influenced by differences in subjective assessment. To test whether the findings were being systematically distorted, in the accompanying meta-evaluation 10 per cent of the sample per evaluation type were read and assessed by at least two individuals. Using statistical methods, the so-called Kappa intercoder reliability coefficient after Cohen was produced. This provides information on the degree of consistency when individual criteria are assessed by two different people. In the accompanying meta-evaluation a Kappa value of 0.63 is achieved, which points to substantial agreement between the individuals involved in assessing the criteria. For a detailed description of the methodology of the meta-evaluation, see Noltze et al. (2018).*

percentages of the various evaluation types. Most of the evaluations are PPRs or KfW ex-post-evaluations. By contrast PEs, and GIZ ex-post and final evaluations, are present in smaller numbers.

Table 1: Population of evaluated projects and size of sample by type of evaluation

Type of evaluation	Number of projects evaluated	Number of projects evaluated in the sample
GIZ ex-post	56	47
GIZ final	44	38
GIZ PPR	343	174
GIZ PE	110	82
Sub-total	553	341
KfW ex-post	462	172
Total	1,015	513

Source: authors' own graphic.

Notes: The size of the sample is dependent on the size of the population and the variance of the score/the proportion of 'sustainable' and 'unsustainable' projects. For further details please refer to Table 9 in the Annex.

3.

METHODOLOGY

This Chapter specifies the regression model used in the analyses, and operationalises the variables contained in it. It then discusses the limitations of the methodology.

3.1 Empirical strategy

As already mentioned in Section 2.1 , there is a conceptual link between the assessment of sustainability and the assessment of the remaining DAC criteria. The findings of the accompanying meta-evaluation also indicate that the sustainability of a project is assessed using criteria that can also be used to assess the DAC criteria relevance, effectiveness, efficiency and impact of a project (Noltze et al., 2018). This link is illustrated in Figure 4. According to this logic, the effect of a variable on the sustainability score is exerted either directly or indirectly – by influencing the other DAC criteria, which then in turn affect the assessment of sustainability. When modelling the determinants, various options arise for incorporating these links. For instance, the average score for all DAC criteria (excluding sustainability) can also be included in the models as an additional control variable. This enables us to distinguish between the effect of a variable on the sustainability of a project and its effect on the remaining DAC criteria. However, this approach is problematic in that it is not possible to determine any effect on the sustainability score for factors that affect sustainability primarily through the other DAC criteria. To capture both the direct and the indirect effect of a factor, we need to exclude the average score for the DAC criteria from the models. The models presented below are therefore estimated both with and without the average DAC score, thus enabling us to assess whether in addition to direct effects there are also indirect effects.

The factors affecting the sustainability score were analysed using multivariate regression models. Multivariate regression models are used to determine the influence of several explanatory variables on one variable to be explained. In this case the variable to be explained is the sustainability score awarded. Here, any given score can be assigned to specific manifestations of a number of explanatory variables. Taking the interplay between the explanatory variables and the sustainability score across a large number of reports, we can then statistically determine the marginal effect of a particular explanatory variable on the sustainability score. Here we distinguish between the effect size (How strong is the effect of the variable, assuming that the other variables are held constant?) and the statistical level of significance of the calculated effect (What is the probability that the observed result will occur, assuming there is no link?). Since the sustainability score involves an ordinal scale – i.e. a ranking with 1 as the top and 4 as the bottom score – we estimated an ordinal logistic regression model. The general formula for the model is

$$N_i^* = \beta_l \boldsymbol{X}_i + \gamma \boldsymbol{DAC}_i + \varepsilon_i$$

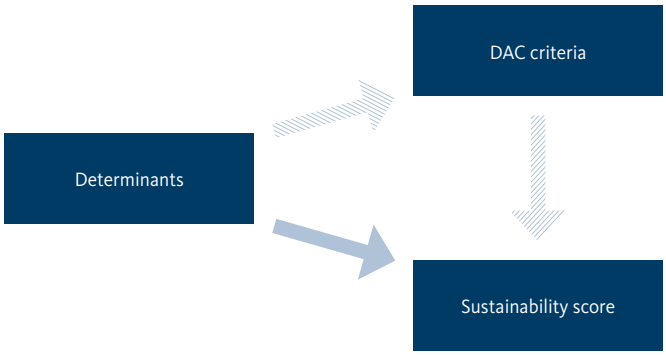
Here, N_i^* represents the latent sustainability score not observed in report i.[22] In the estimated model specification, $\boldsymbol{X}$ is a matrix with explanatory variables. To test how robust the results are, we estimated further models in addition to the model specification described here. These differ in relation to the variables contained in $\boldsymbol{X}$. The modifications are described in Section 3.2 . According to the factors affecting the sustainability rating of specific projects discussed in Section 2.2, $\boldsymbol{X}$ includes specific **characteristics of the project**, **characteristics of the evaluation**, **characteristics of the context in which the project is implemented** and **the criteria used to assess sustainability**. Therefore, the vector β contains the coefficients to be estimated. These specify the effect of the respective explanatory variables on the sustainability score.

$\boldsymbol{X}$ contains certain **characteristics of a project** , namely the duration (years) and financial volume (logarithm of costs in € million), and its overarching development objectives for the social, political, economic and environmental dimensions (number of overarching objectives). The model also specifies which implementing organisation is implementing the project (the GIZ or KfW). Furthermore, it reflects whether there have been delays in implementation of the project (indicator

[22] *The following link exists between the latent variable N_i^* and the sustainability score awarded N_i :*

$$N_i = \begin{cases} 1 \text{ if } Ni* \leq \mu 1 \\ 2 \text{ if } \mu 1 < Ni* \leq \mu 2 \\ 3 \text{ if } \mu 2 < Ni* \leq \mu 3 \\ 4 \text{ if } \mu 3 < Ni* \end{cases}$$

Figure 4: Links between determinants, DAC criteria and sustainability score



Source: Authors' own graphic

Notes: The arrows containing parallel lines represent indirect effects of a factor on the sustainability score. The arrow containing unbroken shading represents direct effects.

variable) and whether a project belongs to the key region for implementation – sub-Saharan Africa – and the key sector for implementation – sustainable economic development (indicator variables).

The **evaluation characteristics** include when the evaluation is carried out relative to the end of the project (years before or years after the end of the project) and the type of evaluation (PPR, PE or final evaluation).

The **implementation context of a project** is modelled by the per capita gross domestic product (GDP) of a country (current figure in US dollars). Here the models also include the Official Development Assistance (ODA) payments received by a country for implementing a project. To guarantee the comparability of transfers received between different countries, the ODA transfers are calculated as a percentage of the country's GDP (ODA/GDP in %). Data on the country's economic development status and ODA transfers are obtained from the World Bank database (World Bank, 2017). The political context of a country is included in the models using the Freedom in the World Index (scale of 1 to 7) published by Freedom House.[23] This provides information on the scope of political rights and civil liberties in a society (Freedom House, 2016). To integrate the aforementioned variables into the models, mean values are calculated for each of the variables for the duration of a project.

The macro indicators used here are aggregated at the level of individual countries. By contrast, development cooperation projects rarely involve the entire territory of a country. They are usually confined to a smaller geographical area. Within a country there may be significant differences in economic, political, social and environmental conditions. These regional differences are not reflected in the existing macro data (Denizer et al., 2013). For instance, a country's average economic growth may be significantly higher than growth in the region of poorest economic performance. In addition to the macro indicators described, the model therefore incorporates the influence of the **project-specific context**. Here the present evaluation synthesis draws on the findings of the accompanying meta-evaluation concerning the criteria used to assess sustainability (Noltze et al., 2018). Based on the review of the stability of the context prescribed in the key questions for assessing sustainability (BMZ, 2006) , the references to the context contained in the report are captured in the model. Here a distinction is drawn between a negative effect of the context on the sustainability of projects, no effect of the context and a positive effect of the context.

In addition to the project-specific context, further criteria for assessing sustainability are also taken from the meta-evaluation (Noltze et al., 2018). As already described in Section 2.2 , the grid for analysing sustainability-related criteria is broken down into seven areas: 1.) Context, 2.) Implementation,

[23] *1 = best score and 7 = worst score.*

3.) Outcome, 4.) Local capacities, 5.) unintended effects (impact), 6.) Predictability of the continuation of results and 7.) Interaction between the dimensions (see Table 6). Systematic analysis of each report using this grid generates a comprehensive picture of the strengths and weaknesses regarding the sustainability of the project. The models also include the effects of these areas on the sustainability of the project. This effect can be 'negative', 'neutral' or 'positive'. Based on the statements made in the report, each of the 48 differentiated criteria is assigned a numerical value. The numerical values for negative (-1), neutral (0) or positive (+1) effects on the sustainability of the project are then aggregated within the 18 criteria/within the seven areas to produce a single value. The more positive (or negative) this value is, the more enabling (or constraining) is the effect of a certain area on the sustainability score.

The vector $\boldsymbol{DAC}$ includes the average score for all DAC criteria with the exception of sustainability. As a control for the results, we estimated a model without the vector $\boldsymbol{DAC}$. ε_{ij} is the normally distributed error term with the anticipated value 0 and constant variance. For a complete list of all explanatory variables including definitions and sources, please refer to the Annex (Table 10).

In addition to the variables described here, there is possibly a link between the methodological quality of reports and the sustainability score awarded. To capture this, the present evaluation synthesis draws on the assessment of report quality performed as part of the accompanying meta-evaluation (see Table 7).[24] The quality assessment of the reports is not included as a control variable in matrix $\boldsymbol{X}$. Instead, it supports analytical weighting for individual observations.[25] On this basis, in the regressions reports of average quality are given a single weighting, reports of above-average quality a larger than single weighting and reports of below-average quality a lower than single weighting.[26] This weighting of individual observations is designed to ensure that the most credible results have the strongest effect in the synthesis. Although the weighting of observations is common practice in quantitative meta-analyses (Borenstein et al., 2009), the methodological quality of reports is not included explicitly as part of the modelling of project performance in any of the comparable studies quoted here (Assefa et al., 2014; Bulman et al., 2015; Denizer et al., 2013; Dollar and Levin, 2005).

The existing cross-section evaluations of German Technical Cooperation do not perform any weighting of observations. Instead, in a number of synthesis studies commissioned by the GIZ the methodological quality of evaluations is used as an exclusion criterion (Caspari, 2014; Huber et al., 2014). In these studies a threshold value for methodological quality is defined a priori, and observations that fall below it are then not included in the evaluation synthesis. Although this approach seems plausible, weighting observations by methodological quality has three key advantages. First of all no arbitrary threshold value is required for this purpose. Secondly, reports are included in the analysis that fall only slightly below a threshold value. Thirdly, the weighting of all observations allows us to differentiate between reports of higher and lower quality.

In addition to the methodological quality of individual reports, we also have to take due account of the fact that the KfW and GIZ use different types of evaluation. For instance, some types of report (KfW ex-post and GIZ ex-post evaluations) assess the sustainability of the project based on actual observations, while other types (PPRs, PEs and final evaluations) base their judgements on assessments of anticipated developments. We are proceeding on the assumption that the two types of report differ systematically with regard to their assessment of sustainability. We therefore estimated the model for two different groups within the sample. Subdividing the observations increases the comparability of the assessments within a group. The first group comprises ex-post evaluations conducted by the KfW and GIZ. The second group contains PPRs, PEs and final evaluations (all conducted by the GIZ).

[24] *For a detailed description of how the methodological quality of reports was recorded and assessed in the meta-evaluation, see Noltze et al. (2018).*
[25] *The methodological quality of the reports is captured as a standardised quality index. This has a mean value = 1 and a standard deviation = 0.5.*
[26] *Due to the fact that the manifestations of the standardised quality index are not integers, the observations are weighted by means of analytical weighting. The weighting is then inversely proportional to the variance of an observation.*

3.2 Sensitivity checks

To test the robustness of the findings, we also estimated further models in addition to the model specifications described. These differ primarily in relation to the explanatory variables contained in $\boldsymbol{X}$. Individual specifications also vary with respect to the number of observations contained in the model – due to the availability of data on particular variables.

Estimating alternative models enables us inter alia to determine whether the findings depend on the operationalisation of particular variables. For example, the political stability of a country and the quality of its institutions can be measured by the World Bank's Rule of Law Index or the Freedom House Index. Furthermore, particular variables can be considered in greater detail. While the main models include for instance only the region where German development cooperation has its main focus, the additional models also assess the effects of all other regions.[27] We also took a similar approach with regard to the sectors.[28] Since regional and sectoral effects may differ between the implementing organisations, we also included interaction terms between implementing organisation and region, and between implementing organisation and sector. The effects of the individual dimensions of the overarching objectives were also analysed in additional model specifications.[29]

Finally, we also looked at whether all the necessary information was included in the model. Furthermore, variables were included in additional models for which not enough information was available for all observations, but which nevertheless might possibly have an effect on the assessment of sustainability. In additional models, we included as characteristics the number of persons involved in the evaluation (number) and the date of the evaluation (year). Other determinants such as the duration of the evaluation (days) or the duration of the field mission (days) ultimately could not be included in any of the models due to the low availability of data. In additional models we did include as characteristics of the implementation context annual economic growth (in %), the World Bank Rule of Law Index (-4 to+4), life expectancy at birth (in years), the population of a country (in millions) and the school enrolment rate (as % of the relevant age group).

As well as projects implemented in a specific country, the sample also includes projects realised in several countries. Indicators aggregated at the country level cannot be assigned to these so-called regional and sector projects. In order to nevertheless include these observations, we estimated additional models that did not include indicators at the country level.

Table 11 (see Annex) contains all variables that were included in the additional models. When presenting and discussing the findings obtained with the main models we will also refer to findings obtained with the supplementary models. The latter do not conflict with the findings from the main models.

3.3 Limitations of the methodology

At a general level we should note that the results from the regression models should be understood as statistical findings that apply across all the evaluation reports included in the analysis. Using the analytical methods applied here it is not possible to study explicitly any specific features of individual projects. This would require a further study that took a systematic look at a limited number of projects– in specific sectors, for instance.

Furthermore, when interpreting the findings we should remember that the explanatory variables in the model may possibly be endogenous. It is conceivable, for instance, that unobserved factors affect certain variables contained in $\boldsymbol{X}$, and at the same time affect the sustainability score awarded. The degree to which the objectives have been achieved at a certain point in the project cycle can for example affect the duration of a project. Projects that are performing well might tend to be extended for that very reason. At the same time project performance can also have a direct effect on project

[27] *The population includes projects in the regions of sub-Saharan Africa, North Africa/Middle East, Asia/Oceania, Europe/Caucasus and Latin America, as well as supra-regional programmes.*
[28] *The sectors of German development cooperation include education, democracy/civil society and public administration, energy, peace-building and crisis prevention, health/family planning/HIV/AIDS, sustainable economic development, food and nutrition security/agriculture/fisheries, transport and communication, drinking water supply/water management/sanitation/solid waste management, environmental policy/protection and sustainable management of natural resources.*
[29] *Another conceivable way to present a project's system of objectives would be to include the number of DAC markers for principal and significant (primary and secondary) objectives. However, these are not included systematically in the meta-data on GIZ evaluations. They are therefore not included in this analysis.*

sustainability. Furthermore, it cannot be ruled out that certain criteria tend to be observed more readily when manifested in a particular way. For instance, negative political conditions may be easier to spot than positive ones. It is conceivable that an evaluator might be more likely to recognise a negative political context if they are already looking at the sustainability of the project in a critical light. In such a case, it is not the political context that is determining sustainability, but the ease with which it can be observed when that particular judgement is being made. In both the above cases, the effect ascertained using the model would be distorted, and hence this would need to be taken into account when interpreting the findings.[30]

Furthermore, the assessment of a project's sustainability may reflect actual sustainability only imprecisely. The assessment process is always subjective. Moreover the accompanying meta-evaluation has demonstrated that the assessment of sustainability in German development cooperation is to a very large extent conducted unsystematically and inconsistently. It is also not clear how the criteria specified in any particular report are weighted when scores are awarded. This sometimes poorly transparent assessment procedure goes hand-in-hand with an allocation of scores (from 1 to 4) which purports to possess an accuracy of measurement that does not exist in this form. To some extent at least, due account is taken of this fact by incorporating the methodological quality of reports into the regression models. The weighting of individual reports allows us to give greater emphasis to links between explanatory variables and the sustainability score in reports that are of above-average methodological quality.[31]

[30] *One way of dealing with the endogenous nature of particular variables can be to use instrumental variables. The instrumental variables are then selected such that they isolate the exogenous variance of the explanatory variables. It is also necessary to ensure that instrumental variables affect the variable to be explained only through the endogenous explanatory variable.*

[31] *An alternative option would be to also incorporate the methodological quality of reports into the model as a control variable. This approach implies that the methodological quality of a report affects the sustainability score only by shifting the y-intercept. However, we cannot rule out that the methodological quality of the report also directly affects the link between explanatory variables and the sustainability score. Weighting the observations by quality of report enables us to represent these links in the model.*

4.

FINDINGS

In this Chapter we will describe the sample first of all with respect to the explanatory variables contained in the model. We will then examine empirically the conceptual link between the DAC criteria of relevance, effectiveness, efficiency and impact, and the criterion of sustainability described in Section 2.1 . After that we will present and discuss the findings from the regression models for each evaluation question. Finally, to obtain the wider findings we will synthesise the individual components.

4.1 Distribution of the explanatory variables by sustainability score

Describing all the explanatory variables contained in the model will make it easier to interpret the regression findings. Table 2 shows their mean values and standard deviations for the sample. The mean values are broken down by the sustainability score awarded. When interpreting the mean values we should remember that differences in individual values between the scores cannot be interpreted as implying a causal link between the variables and the sustainability score. It cannot be ruled out that the variables shown here correlate with other variables that in turn affect the sustainability score.[32]

The findings show that lower average scores for the DAC criteria relevance, effectiveness, efficiency and impact tend to go hand in hand with a lower sustainability score. In addition, we note that shorter projects tend to receive a higher sustainability score. Furthermore, as the volume of funding for a project increases, its sustainability tends to be rated less favourably. Projects in the sub-Saharan Africa region and in the 'sustainable economic development' sector tend to receive lower scores. Within a given score, the percentage of projects in this category increases the lower the score becomes. This is remarkable in that projects in the sub-Saharan Africa region and in the sustainable economic development sector dominate the portfolios of the GIZ and KfW.

Regarding the project implementation context, it emerges that an increase in per capita GDP in a country tends to be associated with a higher sustainability score for a project. By contrast, there is no link between the proportion of ODA transfers received (as a % of national GDP) and the sustainability score for a project. Nor is any link evident between the Rule of Law Index and the sustainability score.

Regarding the criteria for assessing sustainability recorded in the meta-evaluation, it becomes clear that the sustainability score improves the more positively the overall effect of all criteria on sustainability is rated. This pattern is also evident within the seven areas (context, implementation, outcome, local capacities, unintended effects (impact), predictability of the continuation of results and interaction between the dimensions).[33]

[32] *For example, particularly short (or long) projects may be assessed particularly frequently using a certain type of evaluation. The higher scores awarded to shorter projects may be explained by the fact that certain types of evaluation award higher (or lower) scores, and that these types of evaluation at the same time are used particularly often for short (or long) projects. In this case the supposed causal link between duration and score does not exist.*

[33] *The effect of a criterion is either positive, neutral or negative. Consolidating individual criteria into blocks enables us to determine the effect of the various areas on the sustainability score.*

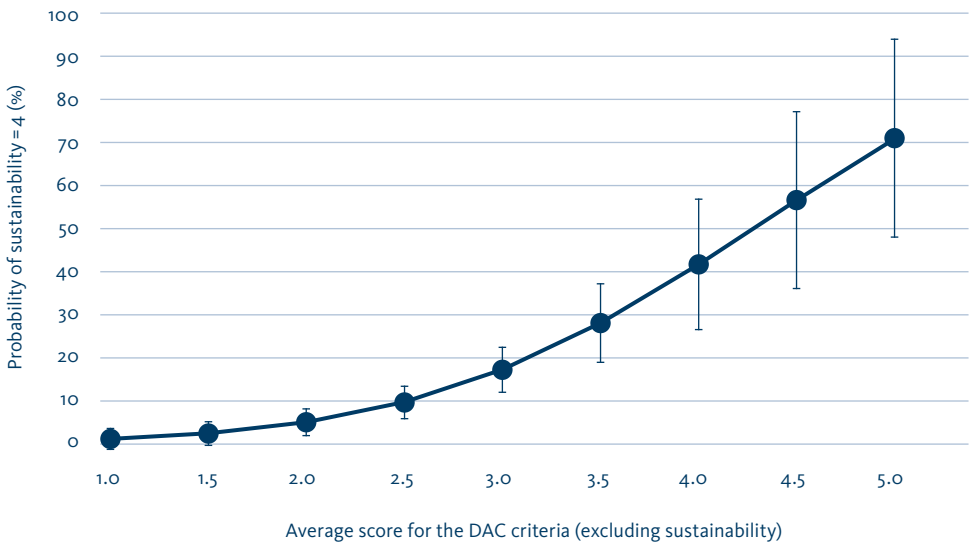
Table 2: Descriptive statistics on the explanatory variables by sustainability score

	Sustainability score			
	1 (n = 30)	**2 (n = 166)**	**3 (n = 256)**	**4 (n = 61)**
Project characteristics				
Score for DAC criteria excluding sustainability (average)	1.8 (0.4)	2.0 (0.5)	2.4 (0.5)	3.2 (0.7)
Duration of project (years)	3.6 (2.1)	4.1 (2.6)	4.6 (3.2)	5.8 (3.9)
Financial volume (€ million) (GIZ n = 297)	7.5 (7.8)	11.3 (15.0)	11.0 (13.9)	11.4 (20.5)
Overarching objective dimensions (number)	1.7 (0.8)	1.7 (0.7)	1.7 (0.7)	1.7 (0.6)
Percentage of projects in the sub-Saharan Africa region (%)	21	30	35	44
Percentage of projects in the sustainable economic development sector (%)	17	25	23	36
Delayed implementation (%)	23	24	34	25
Project implementation context				
Per capita GDP (current figure in US$) (GIZ n = 245, KfW n = 166)	3,203 (2,600)	2,575 (2,671)	2,243 (2,309)	1,868 (1,830)
Net ODA (% of GDP) (GIZ n = 241, KfW n = 165)	5.7 (6.0)	6.0 (7.9)	7.1 (9.0)	6.8 (6.1)
Freedom House Index (GIZ n = 234, KfW n = 158)	4.1 (1.6)	3.8 (1.7)	4.0 (1.5)	4.1 (1.4)
Evaluation characteristics				
Date relative to end of project (years)	0.1 (1.8)	1.0 (2.3)	1.3 (2.5)	3.0 (3.2)
Criteria for rating sustainability (sum of positive and negative effects)				
All criteria	2.9 (3.9)	2.8 (3.3)	-0.1 (3.8)	-4.9 (4.3)
Criteria for context	-0.3 (0.8)	-0.3 (1.1)	-0.6 (0.9)	-1.1 (1.0)
Criteria for planning and implementation	0.5 (1.3)	0.6 (1.0)	0.2 (1.2)	-0.5 (0.9)
Criteria for outcome	2.1 (2.3)	2.1 (1.9)	0.8 (2.1)	-1.2 (2.5)
Criteria for partner capacities	0.7 (1.7)	0.5 (1.7)	-0.5 (1.9)	-2.1 (1.9)
Criteria for unintended effects	0.1 (0.6)	0.2 (0.6)	0.1 (0.5)	-0.1 (0.5)
Criteria for predictability of continuation of results	0.5 (0.5)	0.5 (0.6)	0.1 (0.7)	-0.4 (0.6)
Criteria for dimensionality	0.2 (0.6)	0.3 (0.6)	0.3 (0.7)	<0.1 (0.6)

Source: authors' own graphic.

Notes: The graphic shows mean values and standard deviations for the sample (n=513). This includes 341 observations of the GIZ and 172 observations of the KfW. The figures in parentheses show for how many of the observations information is available on the respective variables. Information on individual variables without parentheses is complete.

Figure 5: Sustainability score as a function of scores for the DAC criteria



Source: Authors' own graphic

Notes: The graphic shows average marginal effects and confidence intervals (95%) for award of the sustainability score 4 by DAC average score. Marginal effects indicate the probability that the sustainability score 4 will be awarded for various average scores for all DAC criteria (excluding sustainability). The findings are based on the main specification of the regression model described in Section 3.1. The model contains 352 observations (KfW ex-post, GIZ ex-post, final evaluations, project progress reviews, project evaluations). The observations are weighted by methodological quality.

4.2 Empirical link between sustainability and other DAC criteria

The conceptual link between the criterion of sustainability and the other DAC criteria was already discussed in Section 2.1. The mean values shown in Table 2 demonstrate that there may also be an empirical link between the sustainability score and the scores for the DAC criteria relevance, effectiveness, efficiency and impact. However, the mean values may also be affected by other variables. Hence from Table 2 we may infer only a correlation between the scores.

Based on the regression model described in Section 3.1 we can determine whether the observed correlations are also statistically significant in the presence of all the variables contained in $\boldsymbol{X}$. Figure 5 shows the effect of the average score for the DAC criteria relevance, effectiveness, efficiency and impact on the sustainability score. The graphic is based on the findings from the regression model. The data points represent marginal effects. These indicate the probability that the sustainability score 4 will be awarded for various average scores for all DAC criteria (excluding sustainability).

4.3 Regression findings

4.3.1 Presentation of the findings

The findings presented here are based on the regression model discussed in Section 3.1. The effect of the sustainability rating criteria is shown both on an aggregate basis across all seven areas (reduced model), and separately for each area (complete model). In addition, the models are assessed both with the average DAC score (excluding sustainability) as a control variable and without the average DAC score. The findings are subdivided into ex-post evaluations (Table 3) and PPRs, PEs and final evaluations (Table 4). Each table then contains the findings for four different model specifications. The regression coefficients for individual explanatory variables are shown. Following the presentation of the findings in an overview, they are then considered in relation to specific variables. The findings are discussed in relation to the evaluation questions.

4.3.2 Effect of project-specific characteristics

To what extent do project-specific factors affect the sustainability score? Figure 6 and Figure 7 show the marginal effects of all project-specific variables contained in the model. Marginal effects are inferred directly from the regression coefficients (see Table 3 and Table 4). They demonstrate the influence that an explanatory variable has on the probability that a certain score is awarded. Marginal effects can be determined for each of the four sustainability scores. It emerges, however, that the majority of projects are awarded the scores 2 or 3 (see Figure 6). This is why we discuss the marginal effects here predominantly in relation to the score 2. However, we do examine the effects in relation to the other scores for all findings.

The findings demonstrate that ex-post evaluations tend to rate the sustainability of projects with a longer duration more favourably. This link is most evident when the score 2 is awarded. Projects lasting around 13 years have the highest probability of obtaining a good sustainability score. As projects become longer (> 13 years) the probability declines, but remains positive overall. When interpreting this effect we should note that the duration of a project may possibly correlate with unobserved factors that in turn affect the sustainability rating. For example, the duration of the project is also a function of the results it has achieved in the past. The more positive the results achieved in the past, the more likely it becomes that a follow-on phase will be approved. At the same time, however, it also becomes more likely that a higher sustainability score will be awarded. In the model for the PPRs, PEs and final evaluations we do not find any effect of the project duration on sustainability rating.

Table 3: Findings from the regression models (ex-post evaluations)

	Reduced model		Complete model	
	with DAC	without DAC	with DAC	without DAC
Project characteristics				
DAC rating excluding sustainability (average score)	2.01*** (0.43)		2.41*** (0.49)	
Duration (years)	-0.01 (0.07)	-0.03 (0.06)	-0.64* (0.33)	-0.49* (0.28)
Duration (years squared)			0.03* (0.02)	0.02* (0.01)
Financial volume (logarithm of costs in € million)	-0.18 (0.19)	-0.24 (0.19)	-0.15 (0.18)	-0.22 (0.42)
Overarching objective dimensions (number)	0.06 (0.43)	0.37 (0.39)	-0.18 (0.43)	0.25 (0.42)
Sub-Saharan Africa (dummy)	-0.07 (0.55)	0.55 (0.52)	-0.14 (0.62)	0.61 (0.56)
Sustainable economic development (dummy)	0.18 (0.62)	0.57 (0.52)	0.24 (0.63)	0.68 (0.53)
Delayed implementation (dummy)	-0.63 (0.46)	-0.47 (0.46)	-0.62 (0.46)	-0.38 (0.46)
GIZ (dummy)	-1.69*** (0.68)	-2.35*** (0.66)	-2.66*** (0.87)	-2.88*** (0.83)
Project implementation context				
Per capita GDP (current figure in US$)	2E-04*** (8E-05)	2E-04*** (8E-05)	3E-04*** (9E-05)	2E-04** (9E-05)
Net ODA (% of GDP)	5E-03 (0,03)	0.02 (0.03)	-0.02 (0.03)	-0.03 (0.03)
Freedom House Index (1–7)	-0.25* (0.13)	-0.27* (0.12)	-0.19 (0.16)	-0.20 (0.13)
Evaluation characteristics				
Date relative to end of project (years)	0.20*** (0.08)	0.22*** (0.08)	0.22* (0.09)	0.23*** (0.09)
Evaluation criteria (sum of positive and negative effects)				
Overall effect	-0.37*** (0.06)	-0.45*** (0.06)		
Criteria for context			-0.09 (0.21)	-0.18 (0.16)
Criteria for implementation			-0.74* (0.32)	-0.61* (0.30)
Criteria for outcome			-0.14 (0.12)	-0.30*** (0.11)
Criteria for local capacities			-0.60*** (0.14)	-0.61*** (0.12)
Criteria for unintended effects			-0.04 (0.39)	-0.20 (0.35)
Criteria for predictability of continuation of results			-0.80* (0.39)	-0.59* (0.34)
Criteria for interaction of dimensions			-0.43 (0.39)	-0.45 (0.32)
Cut 1	-5.95 (3.35)	-11.23 (3.53)	-11.00 (3.95)	-16.47 (3.97)
Cut 2	-1.11 (3.21)	-6.35 (3.27)	-5.91 (3.85)	-11.31 (3.84)
Cut 3	4.25 (3.25)	-1.77 (3.29)	0.66 (3.78)	-6.13 (3.80)
Number of observations	184			
Pseudo R2	0.46	0.39	0.52	0.43

	Reduced model		Complete model	
	with DAC	without DAC	with DAC	without DAC
AIC	246.92	273.18	238.44	270.57
BIC	298.36	321.40	312.38	341.30
Log-likelihood	-107.46	-121.59	-96.22	-113.28
Chi square	96.52	81.36	93.48	110.81

Source: authors' own graphic.

*Notes: Coefficients are shown with the corresponding standard errors. ***, **, * indicate that the coefficients are not equal to zero at a level of significance of 1, 5 or 10 per cent. Levels of significance are based on grouped standard errors at the level of an evaluation report. Cuts 1 to 3 are threshold values that demarcate the individual predicted scores. Pseudo R2 is a pseudo-coefficient of determination of the model whose values lie between 0 (no prediction of the sustainability score) and 1 (perfect prediction of the sustainability score). The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are standards of model quality. The lower their value, the less likely it is that information will be lost. Log-likelihood is based on the sum of probabilities of the predicted and actual findings, and is a model quality standard. The chi square statistic is a model quality standard.*

Table 4: Findings from the regression models (PPRs, PEs and final evaluations)

	Reduced model		Complete model	
	with DAC	without DAC	with DAC	without DAC
Project characteristics				
DAC rating excluding sustainability (average score)	1.35*** (0.46)		1.19*** (0.44)	
Duration (years)	0.16 (0.14)	0.14 (0.14)	0.10 (0.15)	0.09 (0.16)
Financial volume (logarithm of costs in € million)	-0.42* (0.23)	-0.48* (0.23)	-0.52* (0.23)	-0.59*** (0.23)
Number of overarching objective dimensions	0.27 (0.25)	0.25 (0.24)	0.14 (0.26)	0.10 (0.25)
Sub-Saharan Africa (dummy)	0.02 (0.44)	-0.09 (0.43)	0.05 (0.44)	-0.05 (0.44)
Sustainable economic development (dummy)	0.93* (0.50)	0.76* (0.45)	0.91* (0.47)	0.79* (0.43)
Delayed implementation (dummy)	-0.18 (0.44)	0.34 (0.36)	0.02 (0.50)	0.43 (0.43)
Project evaluation (dummy)	0.21 (0.83)	-0.46 (0.76)	-0.20 (0.81)	-0.78 (0.78)
Project progress review (dummy)	0.86 (0.66)	0.25 (0.56)	0.44 (0.68)	-0.12 (0.59)
Project implementation context				
Per capita GDP (current figure in US$)	7E-05 (1E-04)	1E-04 (1E-04)	6E-05 (1E-04)	8E-05 (1E-04)
Net ODA (% of GDP)	5E-03 (0.3)	0.02 (0.03)	2E-03 (0.03)	0.02 (0.03)
Freedom House Index (1–7)	0.11 (0.13)	0.06 (0.13)	0.02 (0.14)	-0.20 (0.13)
Evaluation characteristics				
Date relative to end of project (years)	0.60 (0.47)	0.50 (0.40)	0.46 (0.47)	0.36 (0.40)

	Reduced model		Complete model	
	with DAC	without DAC	with DAC	without DAC
Evaluation criteria (sum of positive and negative effects)				
Overall effect	-0.18*** (0.05)	-0.24*** (0.04)		
Criteria for context			-0.60*** (0.21)	-0.61*** (0.19)
Criteria for implementation			-0.26* (0.16)	-0.31* (0.15)
Criteria for outcome			-0.17* (0.10)	-0.20* (0.10)
Criteria for local capacities			-0.05 (0.08)	-0.11 (0.09)
Criteria for unintended effects			0.54* (0.30)	0.50* (0.30)
Criteria for predictability of continuation of results			-0.76* (0.33)	-0.98*** (0.35)
Criteria for interaction of dimensions			-0.02 (0.26)	-0.03 (0.25)
Cut 1	-4.98 (3.74)	-9.27 (3.56)	-7.96 (3.90)	-12.17 (3.65)
Cut 2	-2.11 (3.74)	-6.54 (3.55)	-5.02 (3.87)	-9.35 (3.60)
Cut 3	1.73 (3.82)	-2.95 (3.56)	-0.86 (3.90)	-5.38 (3.56)
Number of observations	168			
Pseudo R2	0.20	0.16	0.24	0.21
AIC	330.66	343.05	330.25	338.32
BIC	383.77	393.03	402.25	407.05
Log-likelihood	-148.33	-155.52	-142.12	-147.16
Chi-squared	52.50	46.88	65.34	59.28

Source: authors' own graphic

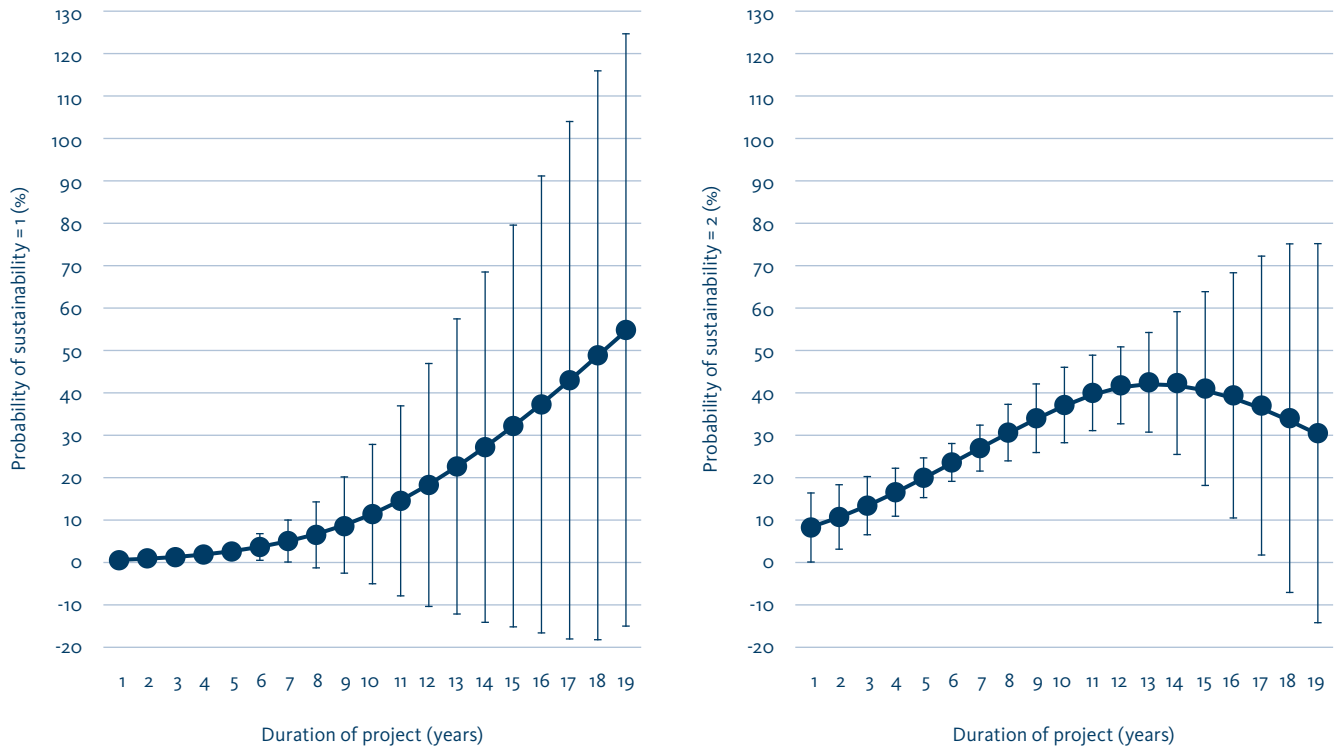
*Notes: Coefficients are shown with the corresponding standard errors. ***, **, * indicate that the coefficients are not equal to zero at a level of significance of 1, 5 or 10 per cent. Levels of significance are based on grouped standard errors at the level of an evaluation report. Cuts 1 to 3 are threshold values that demarcate the individual predicted scores. Pseudo R2 is a pseudo-coefficient of determination of the model whose values lie between 0 (no prediction of the sustainability score) and 1 (perfect prediction of the sustainability score). The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are quality standards of the models. The lower their value, the less likely it is that information will be lost. Log-likelihood is based on the sum of probabilities of the predicted and actual findings, and is a model quality standard. The chi square statistic is a model quality standard.*

Figure 7 shows the marginal effects of further project characteristics. This includes the effects for both ex-post evaluations, and for PPRs, PEs and final evaluations.

In PPRs, PEs and final evaluations, though not in ex-post evaluations, an increase in financial resources for a project is associated with a significantly higher score. In alternating model specifications this effect is not robust.[34] Hence it is not possible to demonstrate a positive link between the financial volume of the project and its sustainability. These findings are consistent with those of empirical analyses of evaluation reports performed by the World Bank (Bulman et al., 2015; Denizer et al., 2013; Dollar and Levin, 2005; Isham et al., 1995). They show that longer and more costly projects do not necessarily lead to improved performance ratings. The findings also demonstrate that an increase in the number of overarching objective dimensions does not affect the sustainability score. Nor does a delay in implementation have any significant effect on the sustainability score. By contrast, as the interval between conduct of the evaluation and the end of the project

[34] *If we exclude macro indicators, the number of observations in the PPR, PE and final evaluation model increases from 168 to 247. The additional observations involve chiefly regional and sector projects. In this model it is not possible to demonstrate any significant effect on the value of a measure on the sustainability score.*

Figure 6: Effect of the duration of a project on the sustainability score in ex-post evaluations



Source: Authors' own graphic

Notes: The graphic shows average marginal effects with the corresponding confidence intervals (95%). Marginal effects indicate the probability that the sustainability score 1 (see graph on left) or the sustainability score 2 (see graph on right), respectively, will be awarded to projects with with a certain duration. The findings are based on the models for ex-post evaluations (see Table 3).

increases, the likelihood of a good sustainability score declines in ex-post evaluations.[35] This finding too is consistent with those of empirical studies (Bulman et al., 2015; Denizer et al., 2013; Dollar and Levin, 2005; Isham et al., 1995).

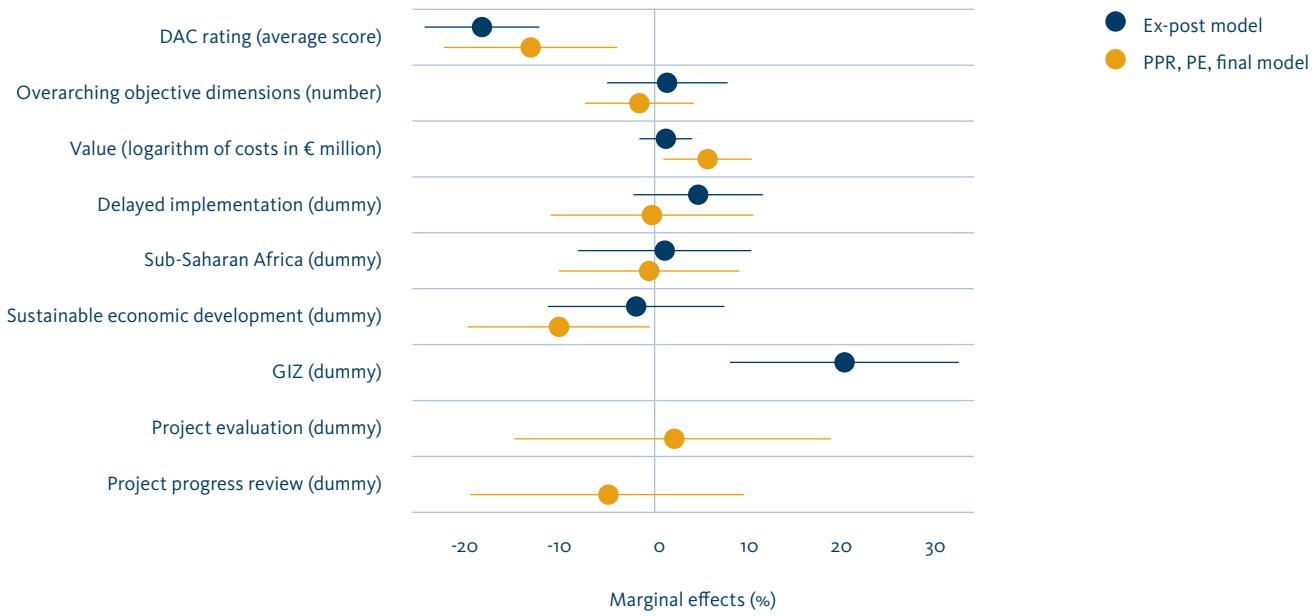
Projects implemented in sub-Saharan Africa are not rated any more or less favourably. Regarding the sector, it is evident that projects implemented in the sustainable economic development sector receive significantly higher scores in PPRs, PEs and final evaluations. This is remarkable in that most GIZ projects are implemented in these sectors, and one might well assume that these projects have comparative advantages over those in other sectors.[36]

The findings also show that in the model for ex-post evaluations there is a significantly higher probability that GIZ projects will receive a higher sustainability score. This point is discussed in more detail in conjunction with the findings on

[35] *Marginal effects are not shown. This conclusion is drawn from the findings shown in Table 3.*

[36] *In addition to the project characteristics presented here, we also tested the effect on the sustainability score of each individual region and each individual sector in which a project was implemented. We also integrated interaction terms between implementing organisation and region, and between implementing organisation and sector, into alternative models. Apart from the links described here we found no further significant effects.*

Figure 7: Effect of project characteristics on the sustainability score



Source: Authors' own graphic

Notes: The graphic shows average marginal effects with the corresponding confidence intervals (95%). Marginal effects show how raising an explanatory variable by one value affects the probability that the sustainability score 2 will be awarded. The findings are shown separately for the ex-post evaluation model, and the PPR, PE and final evaluation model. The findings are based on the complete models (see Table 3 and Table 4). The reference category for the GIZ is KfW projects, while the reference category for PES and PPRs is final evaluations.

the effect of the evaluation criteria on the sustainability score (see Section 4.3.4). In the model for PPRs, PEs and final evaluations there are no significant differences between the evaluation types regarding the award of scores.

4.3.3 Effect of the implementation context

To what extent do context-specific factors affect the sustainability score of development projects? We determined the effect of the national implementation context on a project's sustainability score using several macro indicators. Figure 8 shows the marginal effects of all contextual variables included in the model.

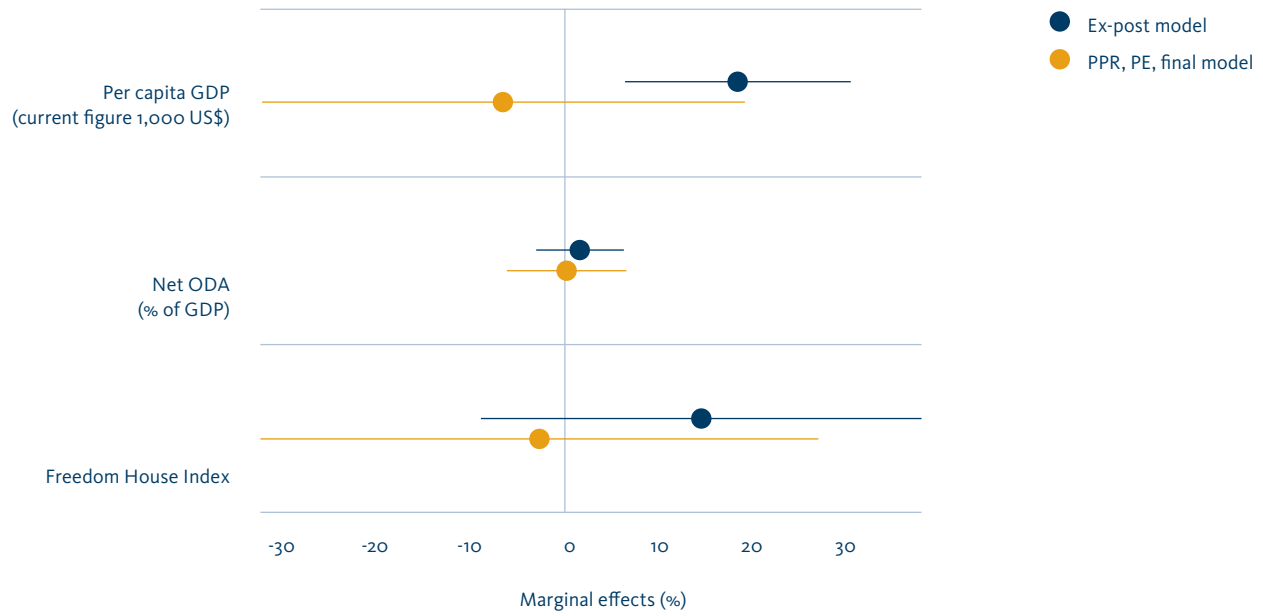
The findings clearly show that in the ex-post model a positive link exists between country's economic development status (measured as per capita GDP) and the sustainability score of projects. They show that an increase of US$ 1,000 in per capita GDP leads to a roughly 2 per cent higher probability that the score 2 will be awarded. Denizer et al. (2013) show that project performance is positively affected by a country's economic development status and its economic stability.

However, we found no link between the national political context (measured using the Freedom House Index) and the sustainability score.[37] This is not consistent with findings in the literature. The latter indicate that a higher degree of rule of law and democracy at the national level are conducive to project performance (Chauvet et al., 2010; Denizer et al., 2013; Dollar and Levin, 2005).

On the other hand, some investigators have found that ODA transfers as a percentage of national GDP can lead to a deterioration in project results (Dollar and Levin, 2005). As ODA transfers increase, partner country capacities can for

[37] *Nor can any significant link be demonstrated when the political context is included in the models through the Rule of Law Index.*

Figure 8: Effect of the implementation context on the sustainability score



Source: Authors' own graphic

Notes: The graphic shows average marginal effects with the corresponding confidence intervals (95%). These show how raising an explanatory variable affects the probability that the sustainability score 2 will be awarded. The findings are shown separately for the ex-post evaluation model, and the PPR, PE and final evaluation model. The findings are based on the complete models (see Table 3 and Table 4).

instance be overstretched (KfW Entwicklungsbank, 2003). We are unable to corroborate this. ODA funding as a percentage of national GDP does not significantly affect the sustainability score awarded. However, here we need to take into account the fact that the allocation of funds may possibly be determined by unobserved factors, which in turn affect the sustainability rating. It cannot be ruled out that ODA transfers are made chiefly to those countries where the need is particularly great and enabling frameworks for project implementation are particularly difficult (Dollar and Levin, 2005).

The fact that the national context appears to have so little effect may be surprising, but this is also corroborated by the findings of Denizer et al. (2013). These findings show that project performance within a country varies more widely than project performance between countries. Hence project-specific factors are more important in explaining project performance.[38] It is possibly the case that, due to their high level of aggregation, the country-level indicators contained in the model do not adequately reflect the immediate context of project implementation.

4.3.4 Effect of the assessment criteria

To what extent do the assessment criteria included in the meta-evaluation affect the sustainability score awarded to projects? The assessment criteria reflect the outputs and results generated by the project. A positive, neutral or negative effect on project sustainability is ascribed to each reported criterion. As explained in Section 3.1 , the criteria are subdivided into a total of seven areas. Within these areas the effects of the individual criteria are aggregated. Positive values indicate that the evaluation judges an area to be

[38] *In addition to the contextual characteristics described here, we also tested how the annual economic growth of a country (%), the World Bank Rule of Law Index, life expectancy at birth (in years), national population size and school enrolment rate affect the score awarded. We did not detect a significant link for any of these factors.*

predominantly conducive to sustainability. Negative values indicate that it sees an area as largely constraining sustainability.

Figure 9 shows the marginal effects of the seven areas of the assessment criteria contained in the model.

These findings clearly demonstrate that certain areas have similar effects on the sustainability score in all types of evaluation. For instance, an increasingly positive rating of implementation leads in both models to a significantly higher probability that a project will be awarded the sustainability score 2 (+6 per cent in the ex-post model, and +3 per cent in the PPR, PE and final evaluation model, if the rating for the area improves by one value). In the area of implementation, the effects of the criteria 'alignment', 'participation' and 'management' on the sustainability of the project are assessed. These criteria are therefore particularly important for rating the sustainability of a project. However, the link identified may also be due to the fact that these criteria are observed particularly frequently when they are manifested positively. In that case it would not be the criteria themselves, but the ease with which their positive manifestation can be identified that is affecting the sustainability score. However, the findings of the meta-evaluation (Noltze et al., 2018) do suggest that there is no one-sided reporting concerning the effects of the criteria in the area of implementation. Also in both models, an increase in positive impressions in the area 'predictability of the continuation of results' raises the probability that a sustainability score of 2 will be awarded (+5 per cent in ex-post evaluations and +8 per cent in PPRs, PEs and final evaluations). The predictability of the continuation of results is a key element of the assessment of sustainability (BMZ, 2006). It therefore comes as little surprise that an important effect on the sustainability score is ascribed to this area. At the same time, however, it is also evident that sustainability is also influenced by factors that go beyond the mere durability of results.

The findings also demonstrate that some areas differ in terms of their effect on the sustainability score depending on the type of evaluation. As the assessment of the project implementation context becomes increasingly positive, the award of a score of 2 becomes more probable only in the model for PPRs, PEs and final evaluations. While the findings of the accompanying meta-evaluation indicate that the context is used to assess the sustainability of the project with particular frequency (Noltze et al., 2018), the regression findings clearly indicate that overall, this more frequent inclusion of the context is only reflected in the score in PPRs, PEs and final evaluations. Given the point in time at which they are implemented, PPRs, PEs and final evaluations rate sustainability above all by assessing future developments. The immediate context of a project is then an important aspect on the basis of which evaluators assess the sustainability of project results.

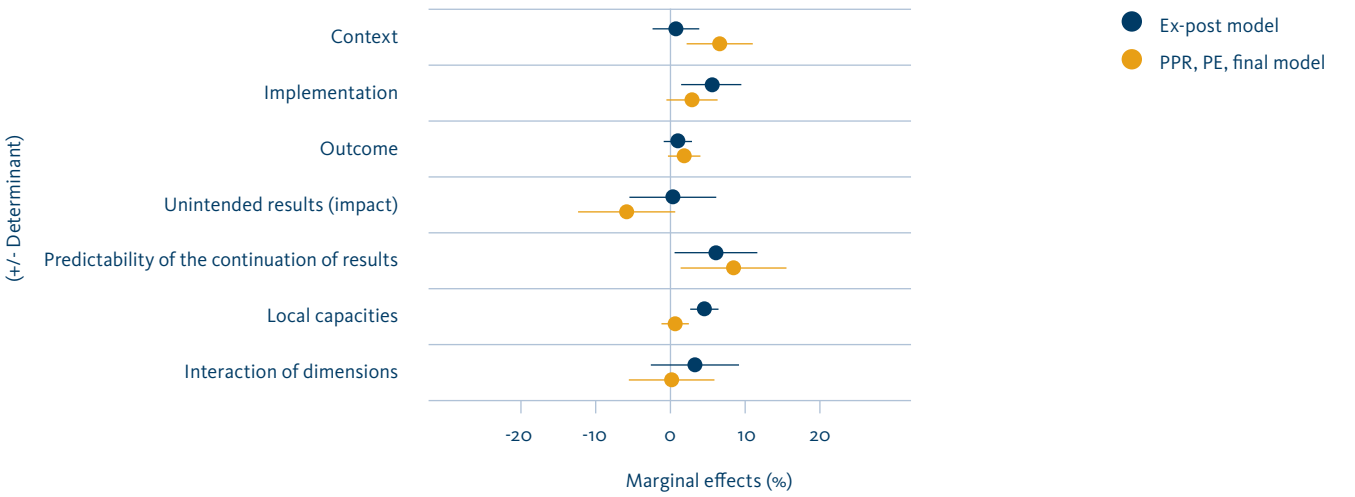
Similarly, it is only in the PPR, PE and final evaluation model that a more positive assessment of the area 'outcome' makes it significantly more likely that a sustainability score of 2 will be awarded (+2 per cent if the assessment of the area improves by one value). Here too the criteria in the area of outcome are used as a basis for the assessment. However, when sustainability is assessed retrospectively – as is the case in ex-post evaluations – outcome plays a more minor role.

The assessment of unintended effects has a slightly significant effect on the sustainability score awarded in PPRs, PEs and final evaluations. The accompanying meta-evaluation shows that this area tends to be included rather infrequently.

In ex-post evaluations, on the other hand, evaluators focus on local capacities. Here there is a significant effect on the sustainability score. A positive assessment of the corresponding criteria increases by 5 per cent the likelihood that the sustainability score 2 will be awarded. Local capacities include financial, technical and institutional partner capacities in the local setting. It seems plausible that these factors will be reflected particularly in the scores awarded in ex-post evaluations. Since ex-post evaluations are conducted several years after the end of a project, the partners are by then solely responsible for implementing and continuing a project, and are therefore probably the focus of the evaluation.

Interaction between the dimensions did not have a significant effect on the score awarded in any of the models. Generally speaking this area is rarely included in the assessment of sustainability (Noltze et al., 2018).

Figure 9: Effect of the assessment criteria on the sustainability score



Source: Authors' own graphic

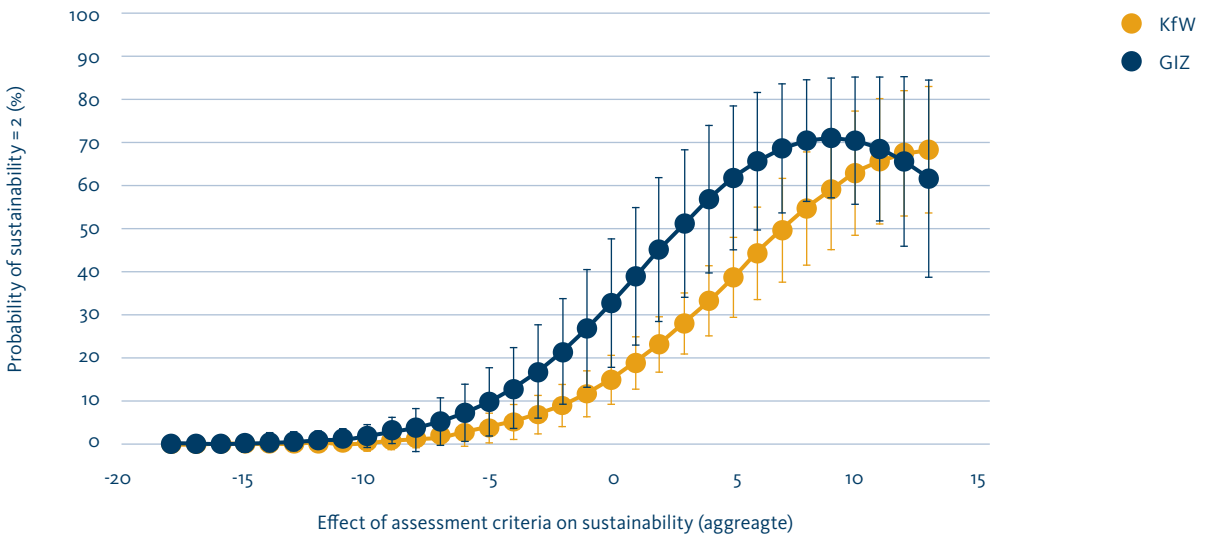
Notes: The graphic shows average marginal effects with the corresponding confidence intervals (95%). These show how raising an explanatory variable affects the probability that the sustainability score 2 will be awarded. The findings are shown separately for the ex-post evaluation model, and the PPR, PE and final evaluation model. The findings are based on the complete models (see Table 3 and Table 4).

The findings in Figure 7 demonstrate that in the ex-post model GIZ projects are significantly more likely (+22 per cent) than KfW projects to obtain the sustainability score 2. As illustrated in Figure 10 , this is directly linked to the effect of the assessment criteria on project sustainability. The x-axis shows the overall effect on the sustainability of a project (the aggregate effect of all 7 areas) captured in one report. The negative x-axis (-18 to -1) represents projects with more negatively rated criteria. The positive x-axis (+1 to +17) represents projects with more positively rated criteria. The y-axis indicates the estimated likelihood that a project will be awarded a sustainability score of 2.

The findings demonstrate that compared to evaluations of the KfW with identical values, GIZ evaluations are more likely to receive the score 2. The differences between the implementing organisations in the value range from -3 to +8 are statistically significant. For instance, the likelihood that a GIZ measure with a value of +5 will be rated 2 is around 61 per cent. By contrast, the probability that a KfW project with the same value will receive the score 2 is only around 41 per cent. These findings suggest that in GIZ evaluations, when values for the criteria are positive there is a stronger overall tendency for this to be reflected in positive scores. When values are in the negative range, however, there are no significant differences between the implementing organisations.[39]

[39] *We also found no differences between project progress reviews, project evaluations and final evaluations with regard to positive values for the assessment criteria and the scores awarded.*

Figure 10: Effect of the assessment criteria on the sustainability score by implementing organisation



Source: Authors' own graphic

Notes: The graphic shows average marginal effects and confidence intervals (95%). The findings are based on the main specification of the regression model described in Section 3.1. The model includes all ex-post evaluations (n = 184). The observations are weighted by methodological quality.

4.3.5 Effect of methodological quality

To what extent does the quality of evaluation methods affect the sustainability score? Although all findings are based on observations weighted by methodological quality, we did not explicitly study the direct link between the quality of reports and scores awarded. Figure 11 shows the link between the quality index and the sustainability score (see Noltze et al., 2018).

As is evident from Figure 11 there is no link between the methodological quality of reports and the sustainability score awarded. This means that evaluations of above-average methodological quality do not award higher or lower sustainability scores.

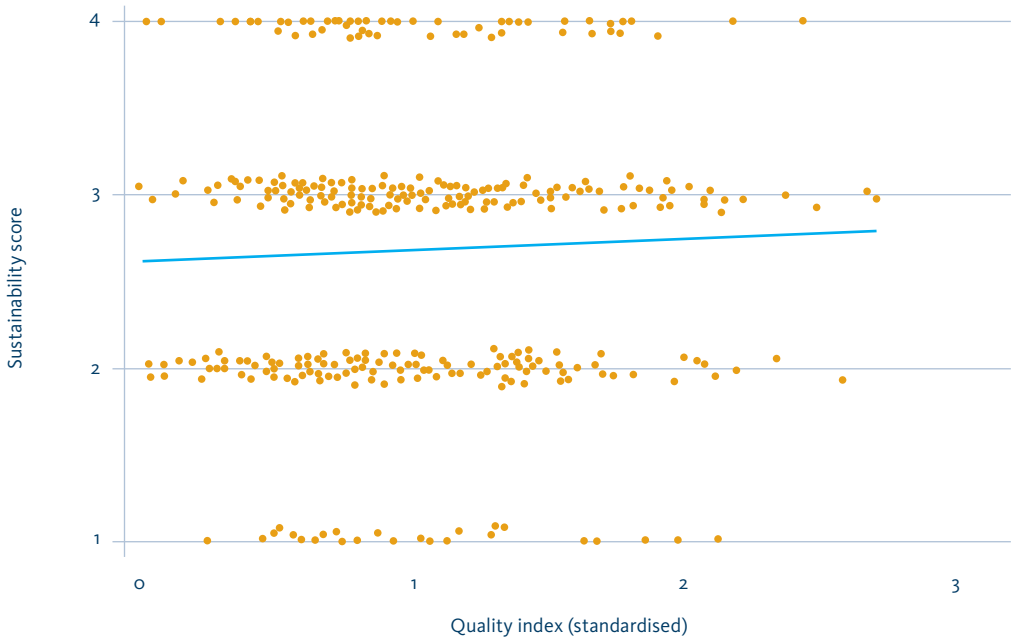
4.3.6 Synthesis

Table 5 summarises the explanatory power of individual variables with regard to the sustainability score awarded. The explanatory variables are shown separately, by average DAC score, project characteristics, characteristics of the implementation context, characteristics of the evaluation and sustainability assessment criteria. These variables are then gradually added to a basic model (without explanatory variables).

The findings show that through the basic model alone, 52 per cent of all scores awarded are predicted correctly.[40] This is to be explained by the fact that in both models (ex-post model, and PPR, PE and final evaluation model), around 52 per cent of all observations obtained a score of 3. When we add to this basic model the average score for the remaining DAC criteria, the proportion of sustainability scores correctly predicted rises to 64 per cent (ex-post model) or 60 per cent (PPR, PE

[40] *The term 'basic model' refers to the model without explanatory variables.*

Figure 11: Effect of methodological quality on the sustainability score



Source: Authors' own graphic

Notes: The graphic shows the standardised values of the quality index for all reports analysed in the accompanying meta-evaluation (n = 513). To make things clearer the data points are shown slightly offset.

and final evaluation model). It also emerges that when the characteristics of the project, the implementation context and the evaluations are added, the explanatory force of both models rises only slightly. This seems plausible, as project characteristics strictly speaking merely smooth the path for sustainable results. They do not have a strong direct effect on the results themselves. If we add the assessment criteria gained in the meta-evaluation by Noltze et al. (2018), the predictive power of the ex-post model improves to 75 per cent. By contrast, in the PPR, PE and final evaluation model there is barely any improvement at all in the prediction of scores. Possibly this is due to the analytical design of the decentralised evaluations. Here the substantiation of results, and thus the substantiation of sustainability, is based exclusively on assessments of the future. True measurement is not possible, due to the point in time at which the evaluations are conducted (which is well before the end of the projects).

Table 5: Percentage of correct predictions and Akaike Information Criterion (AIC) by model specification

	Ex-post models		PPR, PE, final models	
	% of correct predictions	AIC	% of correct predictions	AIC
Basic model	52	405.19	52	376.00
+ average DAC Score excluding sustainability	64	303.64	60	339.91
+ project characteristics	61	305.37	59	340.10
+ characteristics of the implementation context	62	306.83	61	345.00
+ evaluation characteristics	65	305.64	61	346.69
+ sustainability assessment criteria	74	238.44	65	330.25

Source: authors' own graphic

Notes: The model specifications shown here comprise basic models (without expansive variables) that are gradually extended by adding the explanatory variables introduced in Section 3.1. Columns two and four show the scores correctly predicted by the respective model. Columns three and five show the Akaike Information Criterion (AIC). The AIC is a standard of model quality. The lower the value, the less likely it is that information will be lost.

5.

CONCLUSIONS AND RECOMMENDATIONS

Before discussing the findings of this evaluation synthesis, we should once again draw attention to the special features of evaluating sustainability. These are important for understanding the conclusions and recommendations. In evaluations of German development cooperation, sustainability is assessed along with the DAC criteria relevance, effectiveness, efficiency and impact. According to the BMZ's instructions, project sustainability is to be assessed in relation to the continuation of positive results over time, the stability of the context, and the risks and potentials (BMZ, 2006). The specifications for assessing sustainability are conceptually linked to all the other DAC criteria. The findings of the accompanying meta-evaluation demonstrate that this conceptual link is also reflected in evaluation practices. Evaluation practitioners assess sustainability using a large number of different assessment criteria (Noltze et al., 2018). The findings presented here demonstrate that a higher rating of the DAC criteria relevance, effectiveness, efficiency and impact is also associated with a higher sustainability score. The assessment of sustainability thus cannot be viewed in isolation from the other performance criteria. With this in mind, the recommendations made below also apply to areas that can be linked to the other DAC criteria.

The evaluation team will begin by making recommendations on strengthening the sustainability of projects (Section 5.1). This is followed by several overarching recommendations on the comparability of sustainability assessments. These recommendations are designed to foster systematic learning from the findings on projects of official German Technical and Financial Cooperation (Section 5.2).

5.1 Factors affecting the sustainability score

In practice, there are only slight variations in the scores awarded for sustainability in evaluations of German development cooperation. Over 84 per cent of the evaluations studied awarded a score of 2 or 3 for sustainability. By statistical level of significance and effect size, the average score for all DAC criteria (excluding sustainability) is the key determinant of the sustainability score in all regression models. Due to the low variance of the sustainability score and the existence of an explanatory variable of high statistical significance, it is difficult to identify other relevant determinants. Nonetheless, the regression models do demonstrate that certain factors are particularly important with regard to scoring. One explanation for this is provided by the information obtained in the accompanying thematic meta-evaluation. Although the collection of such additional information by means of quantitative content analysis is very complex and expensive, this information does give the evaluation synthesis much greater explanatory power.

The findings demonstrate that in both the ex-post evaluation model and the PPR, PE and final evaluation model, only few factors have a statistically significant effect on the score awarded. Thus the extent to which certain variables play a role in rating depends on when the evaluation is conducted. We will now discuss the main findings and then make recommendations.

5.1.1 Effect of project outputs and results

A synoptic view of the regression models reveals that in addition to the average score for the DAC criteria (excluding sustainability), it is above all the sustainability assessment criteria identified in the meta-evaluation by Noltze et al. (2018) that have a significant influence on the sustainability score awarded. Generally speaking we can conclude that in ex-post evaluations the role and the contributions of development partners and target groups are particularly important for the assessment of the sustainability of projects. By contrast, when sustainability is assessed in PPRs, PEs and final evaluations it is primarily the direct outputs, the implementation of the project and the immediate implementation context that are taken into account. The different weighting of the individual areas is probably due to the different point in time at which the respective types of evaluation are used. While ex-post evaluations base their assessments on observations, in PPRs, PEs and final evaluations sustainability is assessed on the basis of a prognosis. Three to five years after the end of a project it is primarily the partner capacities that can be observed rather than the project implementation structures. If a prognosis is made while the project is still being implemented, evaluators are more likely to base their assessments on the project activities and the immediate context.

However there are also commonalities with regard to the factors identified. For instance, in both ex-post evaluations and in PPRs, PEs and final evaluations, the predictability of the continuation of results has a significant positive effect on the score of project sustainability. This shows that in all types of evaluation the durability of results – a key conceptual element in the assessment of sustainability – has a significant effect on the score awarded.

Regarding project outputs and results, it emerged that sustainability can be increased significantly using the leverage directly available to projects.

Below are our recommendations drawn up on the basis of the findings and conclusions of the evaluation synthesis. These are supplemented in the relevant sub-sections with suggestions and ideas that relate chiefly to their application.

1. The evaluation team recommends that when planning and implementing projects, the BMZ and the implementing organisations should take greater account of the capacities of the local partners and executing agencies, and systematically support their development.
   - With this in mind, an explicit assessment of the capacities of all relevant partners and agencies might also be taken into consideration when deciding on the eligibility for support of a module during project planning. Here it should be ensured that the partners and agencies possess the technical, financial and institutional capacities to continue the activities and outputs previously generated by the project.
   - Furthermore, the capacities of the partners and agencies could be analysed repeatedly at regular intervals in the course of an ongoing project. Successfully transferring the outputs to the partners at the end of the project could also be underpinned by developing long-term exit strategies.
   - Strengthening the partner system might ensure partner-country ownership of implementation of the 2030 Agenda.

2. The evaluation team recommends that the GIZ and KfW in future understand the factors relevant to project management not only in relation to effectiveness, but also in direct relation to sustainability, and take this into account accordingly.
   - These include particularly the use of local institutional structures, the systematic analysis of lessons learned and the development of scaling-up and exit strategies.

5.1.2 Effect of project characteristics

The evaluation synthesis demonstrates that individual project characteristics have a significant effect on the sustainability score. The effect of these characteristics is, however, less than that of a project's outputs and results, hence project characteristics have lower informative value in the models. This is highly plausible because a project's characteristics do not directly affect its sustainability. They rather form the framework for implementation of the project and achievement of its results. As the volume of funding increases, for example, so too does the project's scope for action. However, the effect on sustainability has less to do with the amount of funding and more to do with what the project achieves using its (limited) funds. Nonetheless our findings do permit a number of conclusions.

The core characteristics of the project include its duration and financial volume. Regarding the effectiveness of development cooperation projects, Denizer et al. (2013) establish that longer and more costly projects do not necessarily lead to improved ratings. The findings of this evaluation synthesis are ambivalent in this respect. In ex-post evaluations there is a positive link between the duration of the project and its sustainability. In PPRs, PEs and final evaluations this link is not evident. On the other hand, in PPRs, PEs and final evaluations the financial volume of a project does have a positive effect on its sustainability. In ex-post evaluations there is no such link. The possible effects of duration and financial volume appear rather to be context-specific.

Furthermore, it is remarkable that neither regional nor sectoral expertise have a positive effect on project sustainability. The GIZ and KfW are as 'sustainable' in regions and sectors where they possess a great deal of professional experience as they are in regions and sectors where they are less active.

5.1.3 Effect of the implementation context

The findings of the evaluation synthesis permit us to very largely rule out the possibility that external contextual factors could have a significant effect on sustainability scores. As well as macroeconomic and political indicators at the national level, we also included specific information on the local context of a development cooperation project in the models. According to the regression model findings, neither the national nor the local implementation context of the project has high explanatory power regarding that project's sustainability score. Only the economic development status of a country displays a demonstrably positive effect here, in ex-post models. The low explanatory power of macro indicators at the national level is also evident in similar studies on the effectiveness of development projects, hence it comes as no surprise that this is also the case with regard to project sustainability (Bulman et al., 2015; Denizer et al., 2013). From the point of view of projects themselves this is in the first instance good news, because they are unable to directly influence contextual factors, and more or less have to accept as a fact any effects those factors might have. Sustainability is rather in the hands of the implementing organisations which, together with the partners, executing agencies and target groups in the project country, are responsible for designing and building sustainable structures and processes.

5.2 Systematic, strategic and cross-institutional learning from evaluations

The diversity of assessment criteria, the different types of evaluation, and the variety of formats and content in the compilation of meta-data make it more difficult to compare findings, and thus hinder systematic learning. There are various reasons for this.

First of all, although the key questions for assessing sustainability (BMZ, 2006) do provide guidance on assessing sustainability, they are not sufficiently operationalised. The specific assessment criteria underlying each individual score are many and varied, and cannot be specified unequivocally. The diversity of the portfolio of implemented projects creates a compelling need for flexibility in assessment. Even so, the assessment of sustainability must also be comprehensible and comparable for outsiders. At the turn of the millennium the idea of harmonisation was at the centre of the concept of 'joined-up evaluation'. The 2030 Agenda expresses this idea in the principle of joint accountability.

Secondly, the implementing organisations studied here display systematic differences in assessment practices and evaluation management. The findings demonstrate that GIZ evaluations award significantly higher sustainability scores than KfW evaluations – even though the same number of criteria are rated positively. Furthermore, the use of different types of evaluation leads to structural differences in the assessment of sustainability. Depending on the type of evaluation used, this involves either assessing the future (PPRs, PES and final evaluations) or performing a retrospective assessment (ex-post evaluations by the GIZ and KfW). Furthermore, differences exist in the way the two organisations manage and review their evaluation findings. At the KfW all ex-post evaluations are audited and accepted by the evaluation department. Here the assessment of individual projects is placed in the context of the assessment of comparable projects. Any discrepancies that might arise can then be avoided. By contrast, GIZ's project progress reviews and project evaluations were and are commissioned and accepted on a decentralised basis. Responsibility for conducting the evaluation rests with the officer responsible for the commission for the project in question. Whereas at the KfW a core team of staff members checks all reports, thus establishing a minimum degree of comparability, the decentralised evaluation system at the GIZ precludes the organisation-wide comparison of individual reports. It is therefore to be assumed that overall, evaluations of GIZ projects are more heterogeneous and depend more heavily on attributes of the authors than is the case at the KfW.

Thirdly, the meta-data from evaluations and projects that are recorded by the implementing organisations tally only to a certain extent. Information relevant to the present analysis

was in some cases incomplete, or was systematically recorded by only one implementing organisation.[41]

3. Given the lack of a systematic approach to date in the practice of evaluating and assessing sustainability, as well as aid effectiveness as a whole, the findings of this evaluation synthesis support the recommendation made by the accompanying meta-evaluation to the BMZ that the practice of evaluation by the GIZ and KfW should be harmonised (see Recommendation 8 in Noltze et al., 2018). Several other recommendations concerning the further development of evaluation practices also result. The two recommendations below are also both supplemented with suggestions and ideas that relate chiefly to their application.
   - To guarantee the systematic assessment of sustainability, the evaluation team recommends that the BMZ and the implementing organisations develop standardised and binding criteria. These should serve as a basis for the award of scores, and should be weighted transparently for this purpose.
   - To take due account of the heterogeneous portfolio of German Technical and Financial Cooperation, the criteria should possess an appropriate degree of sector- and region-specific flexibility. Binding instructions on applying the criteria might also be defined separately for each sector or for TC/FC modules.

4. The evaluation team recommends that the BMZ and the implementing organisations – where possible – harmonise the collection of meta-data on projects and their evaluations and record this information at a central point.
   - The systematic and central recording of meta-data from projects and evaluations would make cross-institutional, aggregated analyses considerably easier to perform, and therefore quicker.
   - With this in mind, the BMZ and the implementing organisations might explore how they could meet the requirements of joint accountability articulated in the 2030 Agenda by recording and systematically preparing meta-data.

41 *For example, the OECD-DAC markers for the principal and significant (primary and secondary) objectives of the project are incomplete in the GIZ's meta-data. Neither implementing organisation provides information on the duration of the evaluation (number of working days) or on the time spent in-country by the evaluation mission.*

6.

REFERENCES

Assefa, Y. et al. (2014), Macro and micro determinants of project performance, *African Evaluation Journal*, Vol. 2, No. 1.

Benoit, S. et al. (2017), *Evaluation: a missed opportunity in the SDG's first set of Voluntary National Reviews*, IIED Briefing Paper, IIED, EvalSDG, EvalPartners, London.

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.

Borenstein, M. et al. (2009), Introduction to Meta-Analysis, Wiley, West Sussex, United Kingdom.

Bulman, D. et al. (2015), *Good Countries or Good Projects?*, No. 7245, Policy Research Working Paper, World Bank Group, Washington, DC.

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

Caspari, A. (2014), *Sektorbezogene Querschnittsauswertung: Meta-Evaluierung Ländliche Entwicklung*, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, Eschborn.

Chauvet, L. et al. (2010), *What Explains Aid Project Success in Post-Conflict Situations?*, No. 5418, Policy Research Working Paper, World Bank Group, Washington, DC.

Denizer, C. et al. (2013), Good countries or good projects? Macro and micro correlates of World Bank project performance, *Journal of Development Economics*, Vol. 105, p. 288–302.

Dollar, D. and V. Levin (2005), *Sowing and Reaping: Institutional Quality and Project Outcomes in Developing Countries*, No. 3524, Policy Research Working Papers, World Bank Group, Washington, DC.

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

Hemmer, H.-R. and A. Lorenz (2003), What determines the success or failure of German bilateral financial aid?", *Review of World Economics*, Vol. 139, No. 3, p. 507–549.

Huber, S. et al. (2014), *Querschnittsauswertung Bildung: Meta-Evaluierung und Synthese*, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, Bonn.

Isham, J. et al. (1995), Does participation improve performance? Establishing causality with subjective data, *The World Bank Economic Review*, Vol. 9, No. 2, p. 175–200.

KfW Entwicklungsbank (2003), *FZ-Projekte und Nachhaltigkeit. Zur Berücksichtigung der Nachhaltigkeit durch die KfW in Schlussprüfungen von FZ-Vorhaben: Grundsätzliche Überlegungen*, Nr. 33, Diskussionsbeiträge, KfW Entwicklungsbank, Frankfurt am Main.

Kilby, C. (2013), The political economy of project preparation: An empirical analysis of World Bank projects, *Journal of Development Economics*, Vol. 105, p. 211–225.

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

Lucks, D. et al. (2016), *Counting critically: SDG „follow-up and review" needs interlinked indicators, monitoring and evaluation, IIED Briefing Paper*, IIED, EvalSDG, EvalPartners, London.

Noltze, M. et al. (2018), *Meta-evaluation of sustainability in German development cooperation*, German Institute for Development Evaluation (DEval), Bonn.

OECD (1991), The DAC Principles for Evaluation of Development Assistance, OECD Publishing, Paris.

Ofir, Z. et al. (2016), *Five considerations for national evaluation agendas informed by the SDGs*, IIED Briefing Paper, International Institute for Environment and Development, London.

Schwandt, T. et al. (2016), *Evaluation: a crucial ingredient of SDG success*, IIED Briefing Paper, IIED, EvalSDG, EvalPartners, London.

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.

World Bank (2017), World Development Indicators, *http://data.worldbank.org/*.

7.

ANNEX

7.1 Tables

Table 6: 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 involved in implementing 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.

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 partner 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	This 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	This 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 grid

Notes: For a detailed discussion of the analysis grid, see Noltze et al. (2018).

Table 7: Analysis grid for the assessment of evaluation quality

Areas	No.[42]	Criteria	Definition of the criterion
1. 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.

42 *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.*

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.
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 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.[43]
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 grid.

Notes: For a detailed discussion of the analysis grid, see Noltze et al. (2018).

Table 8: Characteristics of projects, evaluation missions and evaluations by implementing organisation

	GIZ (n = 553)	KfW (n = 462)	% difference
Regional distribution (% of projects)			
Sub-Saharan Africa	29.48	38.74	-31 ***
Asia/Oceania	24.77	25.76	-4ns
Europe/Caucasus	14.65	14.07	4ns
Latin America	13.56	11.04	19ns
North Africa/Middle East	10.31	10.39	<1ns
Supra-regional	7.23	No projects	
Sectoral distribution (% of projects)			
Economy	26.04	19.70	24**
Democracy	23.33	10.39	55***
Water	8.86	18.40	-108***
Health	7.05	14.94	-112***
Environment	12.48	8.44	32**
Other	22.24	28.13	-26**

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

Characteristics of projects			
Start of project (year)	2008 (3.53)	2002 (4.70)	<1***
Duration (years)	3.38 (1.28)	7.25 (3.24)	-114***
Financial volume (€ million) (GIZ n = 473, KfW n = 458)	7.38 (7.31)	42.70 (211.0)	-479***
Markers (number) (GIZ n = 383, KfW n = 434)	2.28 (1.89)	2.65 (1.42)	-16***
Characteristics of evaluations			
Date relative to end of project (years)	0.04 (1.72)	3.41 (2.37)	-8,425***
Field mission (%) (GIZ n = 512, KfW n = 417)	97	79	18***
Evaluators (number) (GIZ n = 537, KfW n = 417)	3.28 (1.37)	3.24 (0.81)	1ns
Sustainability criteria reported (number)	6.19 (0.27)	4.12 (0.28)	33***
Positivity of sustainability criteria	0.33 (0.14)	0.03 (0.13)	91*
Sustainability score	2.55 (0.75)	2.83 (0.72)	-11***
GIZ ex-post	2.75 (0.86)		-3ns
GIZ final evaluation	2.80 (0.63)		-1ns
PPR	2.56 (0.65)		-11***
PE	2.30 (0.92)		-23***

Source: authors' own graphic.

*Notes: The graphic shows mean values and standard deviations for the population by implementing organisation (n=1,015). The values shown in column four indicate percentage differences between the implementing organisations with regard to specific variables. **, *** indicate that the values differ at a level of significance of 5 per cent / 1 per cent. 'ns' means that there are no significant differences. The figures in parentheses show for how many of the observations information is available on the respective variables. Information on individual variables without parentheses is complete.*

Table 9: Sustainability score and scope of sample by evaluation type

Type of evaluation	Number	Sustainability score (standard deviation)	Sample score	'Sustainable' projects (%) (score 1–3)	Sample percentage	Number of observations sample
GIZ ex-post	56	2.75 (0.86)	47	80.4	46	47
GIZ final	44	2.80 (0.63)	34	88.6	38	38
GIZ PPR	343	2.56 (0.65)	110	95.9	174	174
GIZ PE	110	2.30 (0.93)	82	89.0	80	82
Sub-total	553		273	92.4	338	341
KfW ex-post	462	2.83 (0.72)	140	84.2	172	172
Total	1,015		413		509	513

Source: authors' own graphic.

Notes: The size of the sample is determined by the average sustainability score awarded by type of evaluation (sample score) / by the percentage of projects rated 'sustainable' per type of evaluation (sample percentage). The formula used is $sd^2/((\epsilon^2)/(z^2))+sd^2/N$, where sd = standard deviations (sample score) / percentage of 'sustainable' projects (sample percentage), N = population, z = t-distribution value of 1–0.05/2 and ϵ = maximum error. Assumptions: ϵ = 0,1 and z = 1,96 (α = 0.05).

Table 10: Control variables in the main model

Variable	Definition	Unit	Source	Mean value (standard deviation)	
				Ex-post (n = 184)	PPR, PE, final (n = 164)
DAC score	Average rating	Score	Meta-data GIZ and KfW	2.64 (0.68)	2.19 (0.55)
Duration	Length of projects from start to finish	Years	Meta-data GIZ and KfW	6.94 (3.43)	2.77 (1.26)
Financial volume	Total value of projects	Logarithm € million	Meta-data GIZ and KfW	16.03 (1.21)	15.44 (0.95)
Number of overarching objective dimensions	Overarching objectives pursued by the project	Number	Evaluation reports GIZ and KfW	1.46 (0.54)	1.72 (0.79)
Sub-Saharan Africa	Project is implemented in sub-Saharan Africa	Dummy	Meta-data and evaluation reports GIZ and KfW	0.43	0.35
Sustainable economic development	Project is implemented in the sustainable economic development sector	Dummy	Meta-data and evaluation reports GIZ and KfW	0.22	0.29
Delayed implementation	Project is implemented with a delay	Dummy	Evaluation reports GIZ and KfW	0.46	0.19
GIZ	Project is implemented by GIZ	Dummy	Meta-data GIZ and KfW	0	1
Per capita GDP	Per capita gross domestic product	Currently US$ 1,000	World Bank	2.32 (2.58)	2.25 (2.29)
Net ODA	Ratio of Official Development Assistance to gross domestic product	Percentage	World Bank	5.85 (7.31)	5.99 (7.94)
Freedom House Index	Index to capture the status of liberty in a country	Index value	Freedom in the World	3.98 (1.51)	3.90 (1.50)
Timing of evaluation in relation to end of project	Interval between end of project and date of evaluation	Years	Meta-data and evaluation reports GIZ and KfW	3.82 (2.30)	-0.40 (0.64)
Combined effect of criteria	Effect of criteria included in the accompanying meta-evaluation	Aggregate effect of all criteria (-48 to +48)	Evaluation reports GIZ and KfW	-0.22 (5.17)	0.54 (4.19)
Criteria for context	Effect of criteria in the area of context included in the accompanying meta-evaluation	Aggregate effect of criteria (-4 to +4)	Evaluation reports GIZ and KfW	-0.62 (1.13)	-0.5 (1.00)
Criteria for implementation	Effect of criteria in the area of implementation included in the accompanying meta-evaluation	Aggregate effect of criteria (-8 to +8)	Evaluation reports GIZ and KfW	0.11 (0.97)	0.30 (1.29)
Criteria for outcome	Effect of criteria in the area of outcome included in the accompanying meta-evaluation	Aggregate effect of criteria (-13 to +13)	Evaluation reports GIZ and KfW	0.35 (2.54)	1.25 (2.20)
Criteria for local capacities	Effect of criteria in the area of local capacities included in the accompanying meta-evaluation	Aggregate effect of criteria (-9 to +9)	Evaluation reports GIZ and KfW	-0.06 (2.40)	-0.51 (1.79)

Variable	Definition	Unit	Source	Mean value (standard deviation)	
Criteria for unintended effects	Effect of criteria in the area of unintended effects included in the accompanying meta-evaluation	Aggregate effect of criteria (-4 to +4)	Evaluation reports GIZ and KfW	0.14 (0.58)	0.12 (0.63)
Criteria for predictability of continuation of results	Effect of criteria in the area of predictability of continuation of results included in the accompanying meta-evaluation	Aggregate effect of criteria (-4 to +4)	Evaluation reports GIZ and KfW	0.14 (0.68)	0.21 (0.66)
Criteria for interaction of dimensions	Effect of criteria in the area of interaction of dimensions included in the accompanying meta-evaluation	Aggregate effect of criteria (-6 to +6)	Evaluation reports GIZ and KfW	0.30 (0.69)	0.34 (0.76)

Source: authors' own graphic.

Notes: The table shows all control variables included in the main model (see Section 3.1).

Table 11: Control variables of additional models

Variable	Definition	Unit	Source
Regional project	The project is a regional project.	Dummy	Meta-data and evaluation reports GIZ and KfW
Sector project	The project is a sector project.	Dummy	Meta-data and evaluation reports GIZ and KfW
Asia/Oceania	The project is implemented in Asia/Oceania.	Dummy	Meta-data and evaluation reports GIZ and KfW
Europe/Caucasus	The project is implemented in Europe/the Caucasus region.	Dummy	Meta-data and evaluation reports GIZ and KfW
Latin America	The project is implemented in Latin America.	Dummy	Meta-data and evaluation reports GIZ and KfW
North Africa/Middle East	The project is implemented in North Africa/the Middle East.	Dummy	Meta-data and evaluation reports GIZ and KfW
Education	The project is implemented in the education sector.	Dummy	Meta-data and evaluation reports GIZ and KfW
Democracy	The project is implemented in the democracy sector.	Dummy	Meta-data and evaluation reports GIZ and KfW
Energy	The project is implemented in the energy sector.	Dummy	Meta-data and evaluation reports GIZ and KfW
Peace	The project is implemented in the peace sector.	Dummy	Meta-data and evaluation reports GIZ and KfW
Health	The project is implemented in the health sector.	Dummy	Meta-data and evaluation reports GIZ and KfW
Agriculture	The project is implemented in the agricultural sector.	Dummy	Meta-data and evaluation reports GIZ and KfW
Transport	The project is implemented in the transport sector.	Dummy	Meta-data and evaluation reports GIZ and KfW
Water	The project is implemented in the water sector.	Dummy	Meta-data and evaluation reports GIZ and KfW
Environment	The project is implemented in the environmental sector.	Dummy	Meta-data and evaluation reports GIZ and KfW
GIZ and region	GIZ projects are compared with KfW projects in various regions.	Interaction term	Meta-data and evaluation reports GIZ and KfW
GIZ sector	GIZ projects are compared with KfW projects in various sectors.	Interaction term	Meta-data and evaluation reports GIZ and KfW
Overarching economic objective	The project has an overarching objective that is economic.	Dummy	Meta-data and evaluation reports GIZ and KfW
Overarching social objective	The project has an overarching objective that is social.	Dummy	Meta-data and evaluation reports GIZ and KfW
Overarching political objective	The project has an overarching objective that is political.	Dummy	Meta-data and evaluation reports GIZ and KfW
Overarching environmental objective	The project has an overarching objective that is environmental.	Dummy	Meta-data and evaluation reports GIZ and KfW
Rule of Law	Index to capture the level of the rule of law	Index	World Bank

GDP growth	Annual rate of change in gross domestic product	Percentage	World Bank
Life expectancy	Life expectancy at birth	Years	World Bank
Population	Population of a country	Logarithm of population figure	World Bank
Enrolment rate	Primary school enrolment rate	Percentage of children enrolled by age group	World Bank
Number of evaluators	Number of people involved in preparing the evaluation	Number	Meta-data and evaluation reports GIZ and KfW
Date of evaluation	Date on which evaluation was completed	Year	Meta-data and evaluation reports GIZ and KfW
Duration of evaluation	Duration of evaluation from start to finish	Days	Meta-data and evaluation reports GIZ and KfW
Duration of field mission	Duration of field mission	Days	Meta-data and evaluation reports GIZ and KfW

Source: authors' own graphic.

Notes: The table shows all variables used in alternative model specifications (see Section 3.2).

7.2
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

7.3 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
	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