IDOS | German Institute of Development and Sustainability

IDOS DISCUSSION PAPER 14/2023

Learning from KfW's Ex-Post Evaluations?

How Conflicting Objectives Can Limit their Usefulness

Nicola M. Dörrbecker

Learning from KfW's ex-post evaluations?

How conflicting objectives can limit their usefulness

Nicola M. Dörrbecker

Bonn 2023

Nicola M. Dörrbecker is studying for an M.A. in African Culture and Society at the University of Bayreuth, Germany, with elective modules on the Sociology of Africa & Development Policy and on Cultural and Social Anthropology.

Published with financial support from the Federal Ministry for Economic Cooperation and Development (BMZ).

Suggested citation:
Dörrbecker, N. (2023). *Learning from KfW's ex-post evaluations? How conflicting objectives can limit their usefulness* (IDOS Discussion Paper 14/2023). Bonn: German Institute of Development and Sustainability (IDOS). https://doi.org/10.23661/idp14.2023

IDOS Discussion Paper / German Institute of Development and Sustainability (IDOS) gGmbH
ISSN 2751-4439 (Print)
ISSN 2751-4447 (Online)
ISBN 978-3-96021-217-1 (Print)
DOI: https://doi.org/10.23661/idp14.2023

Tulpenfeld 6, 53113 Bonn
Email: publications@idos-research.de
https://www.idos-research.de

Printed on eco-friendly, certified paper.

Preface

The German Institute of Development and Sustainability (IDOS), formerly known as the German Development Institute (DIE), supports a global welfare policy geared towards the concept of sustainable development through interdisciplinary research and impact-oriented policy advice. The research project that I head on “Effectiveness, Knowledge Management and Learning” examines various development organisations in terms of how development bureaucracies manage knowledge and learn from it.

In the context of this research project, I came across the master’s thesis by Nicola Dörrbecker for the degree course on African Culture and Society at the University of Bayreuth. Ms Dörrbecker’s study on ex-post evaluations at KfW Development Bank is, in my opinion, of academic interest due to the quality of the analysis and the practical relevance of the topic. I therefore asked Ms Dörrbecker to prepare an abridged version of the thesis as an IDOS discussion paper.

The discussion paper has undergone an internal peer review process at IDOS, and Ms Dörrbecker has revised the study to address a number of comments. The present discussion paper is a fundamentally revised version of her original master’s thesis. All the interviewees cited here have given their consent for their data to be used.

I am certain that Ms Dörrbecker’s work will make a valuable contribution to the theoretical appraisal and practical implementation of evaluations in German development cooperation. I hope you enjoy reading it.

Bonn, August 2023 Heiner Janus

Acknowledgements

I would like to thank everyone I interviewed for their time and support in answering my many questions.

My thanks also go to Heiner Janus, Daniel Esser, Miriam Amine, Max-Otto Baumann, Stephan Klingebiel and Anna-Katharina Hornidge for their many useful and valuable points and constructive comments. I would also like to thank Katharina Schramm and Alexander Stroh-Steckelberg from the University of Bayreuth, who supervised the master's thesis on which this discussion paper is based.

Bayreuth, 11 August 2023 Nicola M. Dörrbecker

Abstract

The effectiveness of development cooperation (DC) is a topic of extensive debate in this policy field. Yet despite numerous review and evaluation formats designed to promote learning processes and hence enhance effectiveness, it is often impossible to document these improvements. Against this backdrop, the present paper aims to analyse the usefulness of ex-post evaluations (EPEs) by KfW Development Bank – both within KfW Development Bank and at the German Federal Ministry for Economic Cooperation and Development (BMZ), from which it receives its commissions.

Research indicates that EPEs are conducted with great care. Moreover, EPEs can contribute to the legitimacy of (financial) DC, as project results are considered and presented in a structured manner. Nevertheless, the people interviewed for this study regard EPEs as (highly) subjective assessments and believe that these evaluations may under certain circumstances not be comparable with one another. Yet EPEs need to be comparable, because their overall ratings are used to calculate the success rate, which is currently around 81%. This in turn affects KfW's reporting on its performance to BMZ and to the public. The data from the interviews shows that trade-offs during the production and use of EPEs appear to limit the usefulness of this format. EPEs are designed to deliver accountability to the public and to BMZ and to promote learning within KfW. These are conflicting objectives, however, as they would each require a different approach.

According to those interviewed at KfW and BMZ, EPEs are seldom read or used. Interviewees explain that EPEs are rarely relevant to people working in operational areas, as the evaluations are not published until several years after the project concerned has been completed and only occasionally contain information that is relevant to current projects. The evaluations cannot be conducted sooner, however, as otherwise they would not be able to assess the sustainability and development impact of a project. Moreover, interviews and evidence from other studies indicate that EPEs are of limited relevance to political steering at BMZ, even in aggregated form. Nonetheless, the author believes that it would not be an option to no longer conduct EPEs, as they are the only way to review the development impact and sustainability of a representative number of projects in an affordable way, thus forming the basis for delivering accountability.

Reconciling the conflicting goals of learning and accountability is challenging. For the learning component, it would appear to be a good idea to make greater use of cross-sectional analyses and to establish a central support structure for all implementing organisations and BMZ with a view to compiling all the key information from the evaluations and forwarding it to both BMZ and KfW and to the partner countries in a form tailored to meet their needs. For the accountability component, transparency also needs to be enhanced by making completed evaluation reports available to the public promptly and in full. In addition to an evaluation of international research literature, this paper particularly draws on empirical interview data. A total of 13 specifically selected experts from the German DC system were interviewed. This interview data thus forms an illustrative but not representative sample.

Contents

Figures

Abbreviations

BMZ	German Federal Ministry for Economic Cooperation and Development
BRH	Bundesrechnungshof (Germany's supreme audit institution)
DAC	Development Assistance Committee
DC	development cooperation
DEval	German Institute for Development Evaluation
EPE	ex-post evaluation
FC	financial (development) cooperation
GIZ	Gesellschaft für internationale Zusammenarbeit
IDOS	German Institute of Development and Sustainability (formerly known as the German Development Institute, DIE)
KfW	Kreditanstalt für Wiederaufbau (Germany's publicly owned development bank)
OECD	Organisation for Economic Co-operation and Development

1 KfW's ex-post evaluations as influenced by global efforts to enhance effectiveness

The issue of effectiveness is a ubiquitous one in development cooperation (DC). DC is therefore subject to more extensive evaluation than most other policy fields (Faust, 2020, p. 64). Ferguson, Mchombu and Cummings (2008, p. 38) describe this policy field as a real "knowledge industry". The high density of evaluation can be justified from a functional perspective and from the point of view of the legitimacy of this policy field (Stockmann, 2004, p. 3). Historically speaking, the joint evaluation culture created by the member states of the Development Assistance Committee (DAC) of the Organisation for Economic Co-operation and Development (OECD) began at the beginning of the 1990s (Meyer, Bär, Faust, von Jan, Silvestrini and Wein, 2019, p. 165). In 1991, the OECD/DAC Network on Development Evaluation[1] defined five criteria and key questions associated with them as Principles for Evaluation (OECD, 1991). These principles were revised in 2019 and a sixth criterion (coherence) was added (OECD, 2020, p. 1; 2021a, p. 13). The six DAC criteria designed to underpin evaluations are the following: relevance, coherence, effectiveness, efficiency, impact and sustainability (OECD/DAC Network on Development Evaluation, 2019, pp. 7–12).[2]

One tool used to determine results is the ex-post evaluation (EPE). The term refers to an evaluation carried out after a development project has been completed. Its main purpose is to record the development impact and sustainability of a project; to do so, it needs to be conducted some time after the end of the project (German Federal Ministry for Economic Cooperation and Development – BMZ, 2021b, p. 3; Cracknell, 2002, p. 74).

These EPEs are conducted at KfW Development Bank, one of the main actors in German DC.[3] Due to the considerable pressure to justify expenditure in DC (as least as assumed at political level), results have been monitored at KfW Development Bank since 1966 in all the development measures being funded (Terberger, 2011, p. 223). In addition to other formats, around 50% of projects are now subject to an EPE at KfW (KfW, 2021c, pp. 6–7).

KfW Development Bank[4] is part of KfW Group and is responsible for financial cooperation (FC) in development policy. FC is geared towards the development strategies of the partner countries and BMZ's country strategies. The DC programmes, which are aligned with the country strategies, specify the key sectors promoted in the recipient country. An individual project usually contributes to this programme and is set out in detail by KfW. Three actors are generally involved in implementing an FC project. The first is KfW, which uses an operational team consisting of portfolio managers and, where appropriate, technical experts to plan the project, advise the lead executing agency on implementing it and report to BMZ. The second actor is the lead executing agency or project partner in the recipient country – usually a ministry or a state institution. The third actor may be the implementation consultant, who is responsible for technical implementation of the project (KfW, 2022b; Terberger, 2011, p. 228).[5]

1 The Network on Development Evaluation is a working group made up of representatives of evaluation units of the OECD/DAC members' development organisations.

2 For detailed descriptions of the DAC criteria, see BMZ (2020, p. 1) and OECD (2020, p. 2; 2021a, p. 10). A critical discussion of the application of these criteria can be found e.g. in Noltze, Euler and Verspohl (2018b) and in Schönhuth and Jerrentrup (2019, p. 52).

3 In 2020, KfW contributed around EUR 11 billion to development projects, of which some EUR 4 billion came from the federal budget (KfW, 2021b, p. 4); by way of comparison, Germany's entire official development assistance (ODA) flows in 2020 were worth just under EUR 25 billion (BMZ, 2022).

4 Referred to in the following merely as "KfW".

5 In some projects, there is no implementation consultant, while in others the consulting firm has large local offices.

EPEs are used to demonstrate impact and boost the legitimacy of DC, as they consider and present project results in a structured way. However, they also create the impression of being a representative object of FC, since few documents about KfW development projects (other than the EPEs) are available to the public. A project database has been available on the website since the end of 2021 and contains project information on just over 2,100 ongoing KfW projects, but only in a concise form.[6] Overall, the EPEs cover around half of the FC projects. They provide information on objectives, the approach, resources used and partners at the project location and contain a critical analysis of this information. According to KfW, this evaluation system is designed to provide insight on how to improve future FC (KfW, 2022c).

As EPEs are very important for external communication and DC projects are evaluated so extensively, this study poses the following empirical question: How, for what and in what areas are EPEs used in the DC system? This question will be addressed using a theory-based study of the current use of EPEs at KfW. The ex-post evaluation system and specific KfW EPEs are analysed in order to determine their potential benefits. To this end, interviews were held with 15 experts and six long versions of EPEs conducted at KfW were assessed. Moreover, the study examines whether the evaluation purposes proposed by BMZ correspond to the theoretical evaluation functions and looks at how KfW deals with the evaluation purposes and the intended benefit of EPEs. It explores how information may become biased when EPEs are drawn up, how the results are actually processed at KfW and whether the EPEs meet their own standards. The question of how they are used also arises in connection with the political steering of programmes at BMZ, which explicitly aims to improve DC projects by documenting results (Amine & Eulenburg, 2022, p. 1) and to use aggregated findings from project evaluations as a political steering instrument and for knowledge management (BMZ, 2021a, p. 40).

The findings of this study show how conflicting objectives – along with the circumstances under which EPEs are produced – may limit the usefulness of the EPEs, because different functions are often not compatible. The findings are relevant for staff at KfW and at BMZ in terms of what details they can watch out for in future when using EPEs in order to enhance the effectiveness of FC and its capacity for learning. Moreover, this research contributes to a broader academic debate on the topic of learning in DC organisations (cf. for example Dexis Consulting Group, 2020; Hovland, 2003; Kogen, 2018; Krohwinkel-Karlsson, 2007; Yanguas, 2021) and determines how and whether KfW draws lessons from EPEs at an institutional level.

The following section will begin by presenting the methodology of the research on which this analysis is based. This will be followed by the official expectations that BMZ and KfW have of EPEs, the method by which EPEs are produced and the challenges that may be associated with this methodology. The study will then present the various benefits of EPEs determined here and will highlight the lessons that can be learned in this context. Finally, there will be a discussion of the benefits that EPEs might have in future.

2 Methodology and data generation

The question of how EPEs are used in practice is addressed using qualitative methodology. The evaluation process is analysed based on the "following objects" method developed by Czarniawska (2007) from the field of Science and Technology Studies. In this case, the EPE is the "object", the production and use of which is followed. The aim is to determine what functions and what potential use EPEs have within the system and what conflicting goals may limit these

6 The project database can be accessed at: kfw-entwicklungsbank.de/s/dezKK_4. The number of projects it contains was last updated on 31 October 2022.

functions. To triangulate the data on these kinds of conflicting goals, the study looks at how EPEs are interpreted and perceived from the various perspectives of different actors.

The analysis in this study is primarily based on semi-structured interviews with KfW staff. This interview method of qualitative empirical research (after Meuser & Nagel, 1991) is a much-discussed qualitative method in political and social science. The interviews are based on open questions, the analysis of which provides findings and insight to answer the research question. Furthermore, the long versions of six EPEs made available by KfW were analysed. Information from the website (KfW, 2022c) and the synthesis reports published every two years by KfW's FC Evaluation Unit were taken into account. As a result of the COVID-19 pandemic and the associated restrictions, face-to-face talks in person and participatory observation were not possible. The study thus focuses on the aforementioned document analysis and on the 15 semi-structured online interviews.

As the interviewees required a basic understanding of EPEs, they were specifically selected and not randomly chosen. Here, the snowball system (after Kalton & Anderson, 1986; Merton, 1949) was used in which people who had already been interviewed were asked to give internal references. In total, nine KfW and two BMZ staff members, one person from the German Institute for Development Evaluation (DEval) and one from the German Institute for Development and Sustainability (IDOS) were interviewed. The content of these interviews was adapted to each interviewee depending on which division or department they worked in and whether they had already conducted EPEs.

It was not easy to find people willing to be interviewed. Five requests for interviews were rejected at KfW, citing time pressure as the reason, while seven were not answered at all. KfW did not provide access to internal documents, such as an annotated template for EPEs, as the data protection division was not prepared to release them, explaining that a contractual agreement should have been signed with KfW prior to data collection. It was only after the author invoked the German Freedom of Information Act (*Informationsfreiheitsgesetz*) through BMZ that KfW provided access to the long versions of six EPEs requested. In return, the author of this study was asked to sign a data protection agreement concerning the EPEs. When the quotes were submitted for final approval, contacts at KfW said that they expected some of the content of the research work to be modified, both with regard to the quotes by KfW staff and the content of the study as a whole. KfW interviewees requested several feedback loops. In response to another request for consent to the quotes being used in this publication, three interviewees withdrew their consent. The main reason they gave for this decision was their concern that they would have no influence on the interpretation of their (or other) statements. On the whole, data generation at KfW was thus difficult for the author. At the same time, it should be emphasised that several staff members made a great deal of effort and had little sympathy for the unusually complex internal processes.

The research material provides insight into the impact that the EPEs could potentially have at KfW and BMZ. Moreover, it gives an idea of the challenges that may be associated with producing and using EPEs. Due to the way in which interviewees were selected, the material is not representative, however, and does not claim to be exhaustive. There may potentially be other uses not covered here. In addition, the study cannot show whether the challenges associated with the evaluation system are systematically encouraged or facilitated or whether they are individual cases. It can only show the deficits that theoretically exist, but not the extent to which they might be "exploited". As the interviewees primarily came from KfW, there is only limited scope for making statements about other actors involved in EPEs. More interviews would have needed to be conducted in BMZ's country and sector divisions in particular to show how they read and use EPEs. The assessments described in the following sections are thus merely illustrative. Nevertheless, room for improvement in the process design for producing and using EPEs can be identified from the analysis.

The interviews, the long versions of the EPEs, the field diary, email conversations with KfW staff, the KfW Synthesis Report 2019–2020 (KfW, 2021a) and BMZ's Evaluation Criteria for German Bilateral Development Cooperation (2020, in German) were all coded using MAXQDA, a software program for the systematic analysis of qualitative data sets. An open coding[7] mode was used based on Grounded Theory (Glaser & Strauss, 2010). Fifteen interviews were not enough to achieve theoretical saturation, however, nor did the author claim to do so.

Coding was carried out as described by Mayring (2010) using inductive category development and category formation in the process. The first two interviews were thus coded first in order to identify more precise questions for the interviews that followed. These two interviewees were therefore interviewed again at a later stage in order to guarantee comparability. After all the interviews had been completed, generic categories were formed from all the categories and sorted in a mind map to highlight connections between the generic categories. The mind map can be found in Annex 1. After the research had been completed, the findings were made available to all the interviewees to clear up any misunderstandings there may have been with regard to interpreting the answers.

3 Basic theoretical aspects and uses of ex-post evaluations

According to the OECD (1991, p. 5), the term "evaluation" refers to a "systematic and objective assessment of an on-going or completed project, programme or policy, its design, implementation and results." The international literature primarily distinguishes between two different aims of evaluations: accountability and learning (cf. Armytage, 2011; Cracknell, 2002; Kogen, 2018; OECD, 1991; Reinertsen, Bjørkdahl & McNeill, 2022).[8] In German-speaking countries, Faust (2020, p. 64) and Stockmann (2007, p. 37) likewise distinguish between the insight, learning and audit function on the one hand and the accountability function on the other. A different approach has to be taken in the evaluation depending on which goal is primarily being pursued (Stockmann, 2007, p. 37). As accountability and learning are conflicting objectives in EPEs, the methodology needs to be adapted depending on which objective takes priority (Armytage, 2011, p. 267; Stockmann, 2007, p. 38). For the accountability function, the evaluators ought to be external and independent, while for the learning function it is best if they are internal evaluators so that the lessons learned can be put into practice more effectively (Cracknell, 2002, p. 56). Replicability and efficiency are additional requirements along with independency if the emphasis is on the need for accountability. In contrast, if learning is more important, the focus must be on the process. Timeliness is another major factor, as the lessons learned need to be forwarded promptly. It is almost impossible to reconcile these two objectives, and trade-offs mean that neither area is adequately covered (Cracknell, 2002, pp. 55–56; OECD, 2001, pp. 65–67; see also Hoey, 2015, p. 9; Kogen, 2018; Reinertsen et al., 2022, pp. 363–364).

BMZ defines evaluations in DC as systematic and objective analyses and assessments of ongoing or completed development projects. These studies cover the design, implementation and in particular results of the development projects. They are expected to contain practically relevant insights and, where applicable, recommendations on how to improve the design of projects (BMZ, 2021b, p. 3). The term "evaluation" thus includes the application of basic

7 Open coding is an inductive approach in which – without consideration of existing theories – the interview material is analysed and categorised sentence by sentence.

8 Beyond a Eurocentric perspective, there are other positions concerning the objectives that evaluations have in DC; see e.g. Shallwani & Dossa (2023).

principles and the methodologically sound and transparent analysis and assessment of empirical data. The basic principles are the usefulness of the findings (e.g. creation of transparency and practically relevant recommendations), credibility of the findings, independence of the assessment, partnership with the project partners and ethical standards taking account of human rights principles (BMZ, 2021a, pp. 7–8). EPEs are only *one* type of evaluation.

In theory, implementing organisations such as KfW give priority to the learning purpose over the accountability purpose, as they primarily have an interest in enhancing the efficiency and effectiveness of their organisation. The accountability purpose takes precedence for those actors that manage taxpayers' money (e.g. ministries or parliaments) so that they can be held accountable by taxpayers about *whether* cooperation was successful or not. However, they are less commonly interested in *why* cooperation was (not) successful (Cracknell, 2002, pp. 55–56). In practice, too, DC is increasingly focusing on learning. The evidence-based policy-making that draws on this can also be seen in the 2030 Agenda for Sustainable Development, which is predicated on a rigorous results-based management logic (Yanguas, 2021, p. 1). However, it is not clear whether knowledge (which is what evaluations are designed to generate) actually does result in learning and in adjustments being made in the organisation (Dexis Consulting Group, 2020, pp. 13, 33; Hovland, 2003, p. 7; Krohwinkel-Karlsson, 2007, pp. 6–9). Nonetheless, knowledge and learning approaches have been adopted in DC despite the lack of sufficient evidence of their effectiveness (Yanguas, 2021, p. 19). Learning in organisations is often informal, which is why scientific assessment of learning is complex (King & McGrath, 2003, p. 12). Nevertheless, DC has assumed for decades that increased knowledge must lead to increased effectiveness in implementing development projects: by learning from past projects, practices can be adjusted to ensure that resources and approaches are deployed towards better delivery of desired outcomes (Yanguas, 2021, p. 1).

In the interests of accountability, transparency is vital. Disclosing the findings of evaluations ensures transparency concerning the successes and shortcomings of the project, from which consequences can be drawn (Stockmann, 2007, p. 39). There is broad consensus within the OECD that not only the findings but also complete evaluation reports must be disclosed (OECD, 2001, p. 26). Transparency can also generate cross-organisational lessons learned and enables the public and the commissioning party (in this case BMZ) to monitor the effectiveness of DC in a credible way (Borrmann & Stockmann, 2009a, pp. 135–136). For the purpose of boosting legitimacy, data needs to document what input a project uses to generate what output and achieve what outcome/impact what input was required for a project to generate what output and what impact. The funding providers and implementing organisations can thus demonstrate their efficiency and degree of impact. EPEs can also highlight the sustainability of an impact (Stockmann, 2007, pp. 39–40). One of the reasons why evaluations are used is also because of their symbolic effect. This is not their actual purpose, however (Stockmann, 2007, p. 40).

Evaluations are embodied in statute in Germany. The German Federal Budget Code (*Bundeshaushaltsordnung*, BHO) governs the procedure for federal expenditure and revenue. The articles relevant to evaluations are Section 7 BHO – which addresses efficiency and economy, which must be taken into account for all state expenditure and measures – and Section 44 BHO and the associated administrative regulation 11a, which requires progress reviews, including verification of evidence (BMZ, 2021a, pp. 12–13). This is specified in BMZ's guidelines on evaluating DC (first published in 2007, revised in 2021), which cover topics such as reporting and evaluation (BMZ, 2021a). The guidelines are binding for BMZ, the implementing organisations and DEval (BMZ, 2021a, p. 5). Evaluations are only one of the instruments used to review progress. The Guidelines for Bilateral Financial and Technical Cooperation with Cooperation Partners of German Development Cooperation (BMZ, 2021c) also specify that the implementing organisations must carry out their own evaluations based on the OECD/DAC criteria in order to assess the effectiveness of development measures and to derive lessons for

future projects. The reports on final and ex-post evaluations[9] are submitted to the German Government and are also published (BMZ, 2021c, p. 27).

In German DC, EPEs have a range of objectives and functions, which were confirmed by BMZ in 2021 and which are geared towards the DAC Quality Standards for Development Evaluation (OECD, 2010). For BMZ, evaluations are generally a means of learning from experience and being accountable for the results achieved. By carrying out empirically informed analyses and assessments of the success of development measures as objectively as possible, evaluations are designed to help enhance the effectiveness and legitimacy of these measures and to generate evidence for the main users (*insight function*). Evaluations provide a basis for taking steering decisions both on individual projects and on overarching issues in DC (*learning and steering function*) (BMZ, 2021a, pp. 7–8). The evidence generated from the evaluation is intended to contribute to decision-making and to knowledge and data management (BMZ, 2021a, p. 37). The central evaluation units of the implementing organisations (e.g. KfW's FC Evaluation Unit) are expected to focus on final and ex-post evaluations, as an informed assessment of development impact and sustainability cannot be made until after support measures have been phased out (BMZ, 2021a, p. 27).

Due to this consistent results orientation, the sector and policy issues divisions at BMZ assume key tasks in the course of a project cycle:[10] the BMZ divisions are tasked with ensuring right from the start that projects can be evaluated. Every project must therefore have objectives underpinned by indicators at the results level. BMZ's evaluation system relies not only on evaluations by the individual implementing organisations after the projects have been completed, but also on monitoring and evaluation during project implementation (BMZ, 2021a, p. 41). To ensure that evaluation findings can be put to better use, BMZ proposes putting appropriate measures in place to process these findings more effectively and hence translate them into practice with a view to promoting learning processes (BMZ, 2021a, p. 40). The aggregated findings generated by evaluations are also designed to be used by the sector divisions for the future strategic design of DC (BMZ, 2021a, pp. 37–38). Germany's supreme audit institution (Bundesrechnungshof), however, argues that BMZ currently does not manage to aggregate existing evidence and use it for steering or learning purposes (BRH, 2021, p. 20).

KfW clarified the aims of an EPE in 2021, explaining that it had been using EPEs for 20 years to compile knowledge in order to be able to make statements about the impact and sustainability of FC. It added that both positive and negative impacts needed to be examined and quantified with a view to enhancing the effectiveness of FC (KfW, 2021a, pp. 10, 28). KfW also aims to establish the specific impact, hoping to generate knowledge about the successes and impacts of FC. Evaluations are designed to analyse whether projects achieve the goals that have been set for them and to reflect on the background, it says, which it hopes will guarantee the quality of KfW's work and help it learn from experience (KfW, 2022c). According to KfW, this involves both institutional and operational learning, as the lessons learned are designed to be systematically fed back into operational areas (KfW, 2021a, p. 2). Neither KfW's website nor KfW 2021a explicitly mentions that the EPEs, as emphasised by BMZ, are also designed to ensure accountability – nor do they state that the EPEs can be relevant in terms of political steering (KfW, 2022c).

In summary, the goals of the EPEs as defined by BMZ are accountability, learning for future projects, demonstrating sustainability and development impact, institutional quality assurance and (with regard to KfW) analysing the effectiveness of FC. At BMZ, the findings derived from

9 Final evaluations are carried out directly after the end of a project (as is usually the case at the Gesellschaft für internationale Zusammenarbeit, GIZ), whereas EPEs are conducted several years after the end of a project.

10 The project cycle consists of planning, implementation, monitoring, learning and adapting follow-on projects.

(meta-)evaluations[11] are also used for the strategic design of DC. The literature cites goals and functions similar to those set out by BMZ and KfW, and the guidelines are thus theoretically well justified. Some authors doubt whether all the goals of an evaluation can be achieved (at the same time) and highlight conflicting objectives. Germany's supreme audit institution BRH criticises the goals defined by BMZ as being too imprecise and has called for this to be remedied (BRH, 2021, p. 13). In its EPEs, KfW (according to KfW itself) largely focuses on the learning aspect.

4 How are ex-post evaluations produced at KfW? An overview

Having clarified the purpose of an EPE, the process involved in producing an EPE will be described in brief below before assessing the challenges of the methodology in the subsequent section. BMZ has issued precise guidelines on producing EPEs (see BMZ, 2021a, pp. 18–20), on which KfW bases its approach. The main steps will be outlined below. An evaluation process typically involves four steps: preparation and design, data collection and analysis, reporting, and implementation of recommendations (BMZ, 2021a, p. 17).[12]

Preparing an ex-post evaluation

To begin with, KfW decides which projects are to undergo an ex-post evaluation. To do so, KfW's FC Evaluation Unit creates a random sample stratified by sector at the beginning of each year from the projects that have been completed and are ready for evaluation. At KfW, the random sample contains 50% of the statistical population (in 2019/20, it was made up of 149 projects).[13] Projects ready for evaluation are those for which the final review has been carried out and around three to five years have since passed. EPEs are conducted on these projects (KfW, 2021a, p. 42). The random sample guarantees independence and representativeness. A meaningful average rating can thus be determined that is representative for the sector and the year (BRH, 2022, p. 14; Zintl, 2009, p. 247).

Three different groups of people can carry out an EPE at KfW: internal project managers in the FC Evaluation Unit; external evaluation consultants who bid for tenders from the FC Evaluation Unit; and "delegates". The latter are staff from other KfW divisions who agree to carry out an EPE. Of the people conducting EPEs, around 25% are internal project managers, 25% external consultants and 50% delegates. This section describes how delegates conduct an EPE.[14] Evaluators are chosen for the projects to be evaluated on the basis of several criteria: the delegate must not have been involved in the project; to guarantee neutrality and independence, there must not be any direct link to the team or the head of the division; and the delegate should

11 Meta-evaluations are overarching evaluations drawn up from several individual project evaluations in a particular sector, for example; the meta-evaluations assess these evaluations and summarise their findings.

12 For a description of how evaluations are typically conducted, see Stockmann (2007, p. 108) and – specific to KfW – Borrmann and Stockmann (2009b) and Noltze, Euler and Verspohl (2018a, p. 12).

13 The other 50% are not evaluated, so their impact is not examined. Like every KfW project, they are subject to a final review, which is carried out by the operational team after the project has been completed.

14 The internal evaluators from KfW's FC Evaluation Unit go through these steps in a similar way. It can be assumed that external evaluation consultants also take a similar approach to evaluating projects.

speak the local language of the country concerned and have a knowledge of the sector (KFW4;[15] Borrmann & Stockmann, 2009b, p. 250).

While conducting an EPE, delegates receive support from the supervisors in the FC Evaluation Unit on issues concerning methodology, data generation, rating, etc. The supervisors guarantee quality assurance and are responsible for the report (BRH, 2022, p. 4). In this context, "responsible" means creating "objectivity" and comparability in the EPE as, according to KFW2 and KFW4, staff from the FC Evaluation Unit have more experience and are better able to draw comparisons. For example, they know which criteria must not have met for a project to be given a rating of 5[16] (KFW2, KFW4).

Methodology of ex-post evaluations

At KfW, EPEs are conducted using the rapid appraisal method. A description of this method can be found in Beebe (1995) and Chambers (1981). The rapid appraisal approach is based on several (often qualitative) assessment methods and techniques designed to collect data rapidly and systematically if the time on site, the budget or the amount of reliable secondary data is limited. Vindrola-Padros and Johnson (2020) conducted a literature review on "rapid" evaluation methods. In addition to a list of the advantages, e.g. time and cost savings, they also identified limitations, for example the challenge of not achieving the same "depth" or interpretation level as with conventional data analysis methods or a loss of detail and possibly a greater degree of bias on the part of the evaluators (Vindrola-Padros & Johnson, 2020, p. 1600). They point out that these "rapid" techniques are not suitable for all areas of research and that they require methodological adjustments, for which several (experienced) researchers are required (Vindrola-Padros & Johnson, 2020, p. 1601).

At KfW, the rapid appraisal approach is supported or underpinned by quantitative methods (BRH, 2022, p. 12). The evaluators visit the project country and interview the target group, the lead executing agency/partners, other people involved in the project and researchers or sector experts and other development organisations. They consult internal and external information sources,[17] scientific reports and studies and, where applicable, quantitative data sources. Finally, the findings and insights are triangulated (KfW, 2019, pp. 53–57).

The operational team creates a results matrix when the project is being designed. The matrix is a framework specified in connection with the commission concerning how output, outcome and impact will be measured (KFW5); it builds on the results logic (DEV1; GIZ, 2012, p. 28). Inputs are the resources that KfW invests in a project to achieve outputs. Outputs are deliverables created by using the resources. The outcome refers to direct positive and negative results of the project that are generated for the target group by using the outputs (GIZ, 2012, p. 25). The impact describes the development benefit of a project (GIZ, 2012, p. 24). It is not defined at the level of an individual project, but instead at that of the entire DC programme, in other words per country and sector. The individual project must contribute to the DC programme through the project objective and the associated indicators (KFW5; cf. also Amine & Eulenburg, 2022, p. 2). In an EPE, the results logic is regarded as the basis for assessment: the evaluator takes a critical

15 All 13 interviewees agreed to anonymised direct and indirect quotes at the beginning of the primary data collection; however, four of the interviewees revoked their consent prior to this paper being published. To anonymise the remaining interviews, abbreviations of the institutions taken into account in the study were used and the interviewees were numbered from KFW1 to KFW6 and as BMZ1, DEV1 and IDOS1. The job descriptions were given in short form and can be found in Annex 2.

16 The ratings are based on the marking system used in German schools – from 1 (highest) to 6 (lowest), whereby a rating of 4 or below is regarded as "not successful".

17 These include sources such as progress reports, final reviews, consulting reports, reports by the partners, etc.

look at it, uses it to examine whether and to what extent the project has contributed to achieving the objectives and looks at how the indicators from the results matrix have developed over time (DEV1; BMZ, 2020).

To establish without doubt whether an impact has been achieved by the project, a rigorous impact evaluation[18] would need to be conducted (BRH, 2022, p. 13), because EPEs often work on assumptions about the impact that the project *might* have. According to KfW, rigorous impact evaluations are not worthwhile for "normal" EPEs due to the high costs and the relatively limited benefit. A study by Krämer et al. shows that the usefulness of rigorous impact evaluations is often much higher than its costs, however (Krämer, Jechel, Kretschmer & Schneider, 2021, p. 36). Although the control group method is not used in EPEs, control group comparisons can be carried out, e.g. using satellite data, without having to collect data in the control group right from the start (KfW, 2021a, pp. 28, 34) as would be necessary for a rigorous impact evaluation.

Data collection for ex-post evaluations

Data collection usually involves a trip to the project country lasting around two weeks.[19] Prior to this, a list of interview questions is drawn up and individuals to be interviewed are identified (KFW3, KFW4). As the project regions and the target group of the projects are often large, a random sample is generated of the areas (villages, where appropriate) and of the representatives of the target groups.

For each DAC criterion, certain key questions are examined relating to the principles of the 2030 Agenda, among other things (DEV1). They are stipulated in the evaluation design by the guideline on dealing with the OECD/DAC evaluation criteria (BMZ, 2020) and must be taken into account both in designing new projects and in producing each report during the project term. The key questions are designed to ensure comparability.

The ratings are awarded by the evaluators on the basis of quantitative data and qualitative impressions from the interviews. A project's overall rating is not simply the average of the ratings for the individual DAC criteria (KfW, 2021a, p. 43); instead, there are three essential criteria: effectiveness, development impact and sustainability. If there is a serious shortcoming in one of these three areas, it cannot be offset by a positive assessment of other criteria and the entire project can no longer be rated as "successful overall" (KfW, 2021a, p. 42). The project is thus given an overall rating of less than 3 (DEV1).

The draft report is sent to the project managers or to the relevant operational department. They are given the opportunity to comment if they believe that facts have not been presented accurately or are missing. According to the people interviewed for this study, this does not usually result in any major changes, not even in the rating (KFW2, KFW5, KFW6). Finally, the completed EPE is sent via the evaluation mailing list to the team or division responsible for the project and to BMZ, among other addressees, and is stored in the document management system. Part 1[20] is published on the KfW website and sent to the partner/lead executing agency. Part 2 is not available to either the partner or the public.

18 Rigorous impact evaluations can attribute results/effects *causally* to a development project. A precise explanation is given in KfW (2022a).

19 In some cases, desk/remote evaluations are carried out.

20 Part 1 of the EPE begins with a cover sheet showing the data, objectives and implementation of the project, a list of key findings in brief, the overall assessment in text form and the conclusions. From p. 1 onwards, the individual DAC criteria are listed with text and ratings. Part 1 is around six to eight pages long and is available on the KfW website in German and English. Part 2 covers aspects such as methodology and contains more detailed explanations of the statements made in Part 1.

5 Challenges in ex-post evaluations: what information does the interview data provide?

In theory, the processes and workflows involved in conducting an EPE would appear to be reasonable and are logically derived from the rapid appraisal method. In practice, EPEs involve considerable time and work and are conducted with great care, although experts familiar with the system say that they have limitations. An analysis of the data obtained from the interviews shows potential challenges in implementing the methodology of the EPEs and the process. An overview of the main criticisms of the methodology mentioned by the interviewees will be presented below, underpinned by theoretical considerations. These points of criticism were derived inductively from the interview data and summarised in generic categories, which will be described separately below. As the sample was not representative, the analysis of these criticisms does not claim to be exhaustive. However, the data particularly illustrates challenges that are repeatedly mentioned. These views are put in a theoretical framework to highlight options for making improvements.

Possible biases in the methodology and data generation

All EPEs are assessed by rapid appraisal using the same method. A distinction is not always made in terms of the size and extent of the project.[21] No counterfactuals are construed, in other words alternative scenarios for using the funding (IDOS1; cf. White, 2009, p. 213). The quality of a rapid appraisal method depends on the complexity of the question being asked: the more complex it is, the more difficult it is to answer within a short space of time. Whether the rapid appraisal method is adequate is thus heavily dependent on the type of project (KFW4). For a statistically representative random sample, there is not always sufficient time for an EPE: KFW4 states that evaluators cannot conduct more than ten to 12 interviews a day. If the evaluation trip lasts for a maximum of 12 days, the evaluator cannot interview a representative sample if the target group is made up of several thousand people and the area is also very large (KFW4). In order to minimise this problem, KfW tries to use local appraisers, if possible (KFW4). DEV1 believes that the limited number of evaluation days at the project location are sufficient to answer certain questions; on the other hand, based on the DAC criteria, questions are asked about the project that the evaluator might not be able to fully answer during the limited time available. Rapid appraisal is an appropriate format only if KfW's monitoring systems are comprehensive. At present, however, the monitoring systems tend to be fairly poor (DEV1; cf. Hartmann, Amine, Klier & Vorwerk, 2019, p. 48). KFW2 was surprised how extensive and careful evaluations were when the rapid appraisal method was used, however.

In some cases, the lead executing agency itself selects the target group representatives to be interviewed or the villages and regions that the evaluators are to visit. This can lead to a bias in the findings if the evaluators uncritically adopt the sample chosen or have not received adequate training in this area. Nevertheless, evaluators can distance themselves from the partner's suggestion and select locations as they see fit (KFW4). Sometimes the evaluators ask the operational teams to recommend project locations (KFW3). In some cases, lead executing agency staff accompany the evaluators to the project locations or regional KfW staff accompany them to talks with the lead executing agency (KFW2). As a result, interviewees might be intimidated or feel that they are being monitored and might adjust the answers they give accordingly (KFW4, IDOS1). Data may also be distorted if target group representatives who live near a road are interviewed while those who live further away are not, for example (KFW4), because projects and their EPEs are implemented in countries in which the infrastructure is

21 Rapid appraisal may be supplemented by quantitative methods, e.g. satellite data analyses.

sometimes not well developed and the evaluators can therefore only visit some of the project locations. The country context must thus be taken seriously as a source of bias (IDOS1).

Challenges relating to indicators and to formulating objectives

Imprecise and poorly formulated objectives are another problem and may mean that an evaluation is automatically positive (Wilhelm, 2015, pp. 10–11). Objectives are either formulated vaguely right from the start or the targets are too low (Amine & Eulenburg, 2022, p. 3; Stockmann, 2007, p. 64). An inadequate data basis is another challenge, as data is sometimes difficult to obtain in "developing countries" (Wilhelm, 2015, pp. 10–11).[22] The interviewees say that people sometimes "forget" to obtain baseline data at the beginning of the project. The data collected for the final review or the EPE therefore cannot be compared with the baseline data to determine what developments have taken place. In this area, it might be a good idea to establish closer cooperation with the partner (Holzapfel & Römling, 2020, p. 35). Although many indicators[23] can show outputs and possibly outcomes, they do not allow any statements to be made about the broad impact of a project (Borrmann & Stockmann, 2009b, p. 256; Holzapfel & Römling, 2020, p. 31). In this study, too, problems were seen with indicators: a project is ultimately judged, among other things, by whether it fulfils the indicators set out at the beginning. In the biodiversity sector, however, there are no suitable indicators to measure the impact in some cases or the impacts might not have been fully realised by the time they are measured, for example (KFW5). Several interviewees described this problem in a similar way.

The delegates system as a possible source of bias

According to Germany's supreme audit institution BRH, both BMZ and KfW have emphasised that the EPEs conducted by delegates meet the quality standards as a result of supervision by the FC Evaluation Unit and that the delegates meet the requirements (BRH, 2022, p. 10). On the other hand, BRH clearly states that the FC Evaluation Unit cannot fully offset serious methodological deficits on the part of evaluators [referring to delegates] without an unreasonable amount of effort (BRH, 2022, p. 11).

The evaluators who produce an EPE need *comprehensive knowledge* of countries, sectors and project types. Even if they have this knowledge, it is difficult to transfer it to an EPE (BMZ1). Delegates often evaluate sectors and countries in which they do *not* work (KFW6), which means that they may lack background knowledge about the country and sector. The evaluator also needs to consider whether or not countries can "put on a perfect Potemkin display" (KFW4). In some cases, KfW operates in countries in which access to local actors is difficult or in which these actors cannot talk freely (IDOS1). If the evaluators cannot assess these conditions (accurately), this may lead to biases.

Delegates are more familiar with the processes and workflows in FC and at KfW than external evaluators are and are therefore better able to assess them, but they may have methodological deficits. However, external evaluation consultants are not automatically a quality criterion (KFW2), as they may give their assessments a more positive spin in order to receive follow-on

22 This generalisation needs to be used with care: many "developing countries" have made extensive progress in recent years in developing their national evaluation systems and are therefore now able to aggregate data themselves and make it available. However, it would appear that the partners' systems have rarely been worked with to date.

23 Indicators are a broad topic – not only in German DC – and cannot be comprehensively covered in this paper. Literature on indicators used by KfW can be found in KfW 2021c; on the discussion about indicators (in DC), see e.g. Armytage (2011, pp. 267–268), Bartl, Papilloud & Terracher-Lipinki (2019), Goodwin (2017), Rottenburg, Merry, Park & Mugler (2015) and Sabbi & Stroh (2020).

commissions (KFW4). According to the interviewees and the literature, the conclusion to be drawn is therefore that as far as bias is concerned, neither delegates nor evaluation consultants are a perfect solution. A combination of delegates, evaluation consultants and FC Evaluation Unit staff would thus appear to be the right approach, although delegates should possibly receive better training.

Objective and subjective assessment in ex-post evaluations

EPEs should be as objective as possible.[24] Objectivity is understood as meaning that independent researchers using the same instruments to conduct a study obtain the same statistical findings. This guarantees the comparability that BMZ requires. According to BMZ, this should be ensured using evaluation questions adapted to the issues being addressed and uniform evaluation standards in the form of ratings (BMZ, 2021a, p. 28). Nevertheless, the findings of an evaluation are dependent on observation. Evaluators adopt a particular viewpoint and therefore cannot be completely independent or objective in their assessment (Wilhelm, 2015, pp. 59–60; on the limited extent to which objectivity can be achieved in science, see also Weber, 1977). Many interviewees mention subjectivity in the EPEs: KFW2, KFW4, KFW6, BMZ1 and IDOS1 explicitly consider EPEs to be (very) subjective, as they are based on the experience of the evaluators, who can determine their own focus and have a degree of discretion. However, "objectivity" and derivation are mandatory for all evaluators under the DAC criteria (KFW2). In contrast, KFW5 remarks that the DAC criteria can be met very individually by different evaluators, adding that they attempt to fulfil objective scientific criteria, but that EPEs are not actually objective (IDOS1). However, if the findings suggest other conclusions or assessments, this also needs to be outlined (KFW4).

Subjectivity not only concerns the interpretation of the answers but also how interviewees actually answer – partly due to their own subjective views and partly for strategic reasons. It is thus impossible to achieve complete objectivity, even though the FC Evaluation Unit attempts to do so. Nevertheless, the EPEs must be comparable, as the data otherwise cannot be used to determine the success rate, for example.

Rating system in ex-post evaluations: objective or subjective?

Another major part of the EPEs is the rating system. The evaluators award a rating for each DAC criterion in the EPE and these are used to calculate the overall rating. This is designed to translate the statements made in the text into an assessment based on transparent standards. The rating should not be viewed in isolation, however, but instead in connection with the underlying text on which the rating is based. For KFW2 and KFW4, this text is more important than the rating itself. Even though two interviewees share this view, particularly good or poor ratings are nevertheless highlighted and discussed.

According to Porter, ratings are always subjective to some extent (Porter, 2015, p. 36). Objectivity is seen as being almost impossible to achieve, and even standardised lists of questions cannot completely eliminate subjectivity. KFW4 believes that even formalisation does not make an assessment more credible. Ratings are awarded not only on the basis of "common sense" and experience but also as the evaluator sees fit, offering "room for discretion" (KFW4). KFW6 also sees ratings as subjective, as the facts are not always unambiguous. According to DEV1, evaluators ask all the interviewees the same questions and, like all their colleagues, collect and analyse data regarding the same facts. Nevertheless, they may arrive at different assessments, particularly in the case of delegates who have little experience of EPEs. The FC Evaluation Unit tries to compensate for this by also comparing the ratings in the draft report with

24 Objectivity is one of the quality criteria of empirical social research (see Lienert & Raatz, 1998).

EPEs carried out previously (during the past ten to 20 years) or at the same time on similar projects and examining whether a particular EPE is an outlier – in other word, whether projects are systematically rated more positively or negatively than during the past 20 years. This comparison is discussed within the FC Evaluation Unit and adjustments may be made if there is any doubt (DEV1). This can lead to a certain degree of path dependency, however.

Rating system of ex-post evaluations: incentives and biases

Towards the end of the process of producing an EPE, the evaluation is sent to the operational team of the project concerned so that they can provide feedback. Interviewees mention (frequent) discussions between evaluators and the operational staff during these feedback processes (KFW2, KFW4, KFW5). For some projects, the rating is irrelevant for the operational staff, however, and there are no discussions, e.g. in the case of concepts that are "non-sellers": "The operational staff do not care as long as the rating isn't a 4 or worse. If a non-seller is given a 4 or worse, people wonder whether they might be right" (KFW4). KFW4 also says that a rating of 2 or 3 attracts little interest. As described in Section 6.3, these ratings are the most common ones, however.

As already emphasised, the text generates less interest than the ratings. KFW3 and KFW6 say that there is a certain pride attached to ratings. As soon as numbers are being used, the products with which they are linked are objectivised and neutralised and tend to become a new "reality" (Desrosières, 2015, p. 334). Ratings are easier for staff to take up and process. According to interviewees who have worked on projects for longer, a good rating in the EPE is a source of pride. This suggests that EPEs are taken seriously (which points to the independence and accuracy of the assessment) and that staff themselves are motivated. However, this can also lead to positive biases in the rating, as those working on the EPE have an interest in rating the project more positively than it is for the reasons outlined above. The discussion about ratings during the feedback process for EPEs indicates that they do in fact play an important role. In addition, interviewees state that they do not read or question the text below good ratings; this, too, suggests a degree of pride. If it was just about learning, people ought to be looking at good ratings too in order to identify any mistakes.

The political setting can also have a distorting effect on evaluations (Henry & Mark, 2003, pp. 301–302) and on the rating given. One example here is a meta-evaluation by Kirsch and Wilson about EPEs by KfW and GIZ from Afghanistan: due to the considerable international interest and the controversy surrounding the military mission, there was greater political pressure to achieve quick results in DC there (Kirsch & Wilson, 2014, p. 27). The information from the evaluations was used to confirm political decisions that had already been taken instead of highlighting what worked (or did not) at local level. None of the projects was given a rating worse than 3, even though the outcome and impact were barely measured (Kirsch & Wilson, 2014, pp. 33–34). BMZ's official statement on the meta-evaluation by Kirsch and Wilson was also critical of the fact that they were not adequately measured (BMZ, 2014). This example shows how ratings may also be influenced by external circumstances and political interests.

6 Ex-post evaluations and their impact: how are findings used?

6.1 Who reads the ex-post evaluation?

In order to understand what impact EPEs have, we need to determine who reads the reports, what is read and why (not), and how evaluation findings as a whole are used, because lessons are not learned from evaluations solely by reading the reports. Every year, countless documents are produced on the topic of DC. However, this knowledge cannot have an impact if it goes unread (Yanguas, 2021, p. 9) or is barely used in formal processes (Krämer et al., 2021, p. 38).

The *primary target group* of the EPEs are the BMZ officers (DEV1). The interviews from this research suggest that EPEs are only read to a limited extent at BMZ. Although BMZ insists that all DC organisations must report on the projects (BMZ, 2021a, p. 12), this leads to extensive reporting to the BMZ divisions, which cannot read all the reports in their entirety (presumably due to a lack of time). The EPEs thus compete with a large number of other reports and activities by the country divisions (KFW2; Hartmann et al., 2019, p. 7).

At KfW, the EPEs (Parts 1 and 2) are available to all staff within the organisation (Borrmann & Stockmann, 2009b, pp. 262–263). What and how operational staff read does not differ noticeably from the situation at BMZ: the interviewees primarily read the EPEs on their own projects, but often only the cover sheet and at most Part 1. If their own projects are evaluated, the operational team (if it still exists) reads the EPE due to the feedback process (KFW2). KFW6 reads evaluations on his own projects, in some cases during a follow-on phase to check on something from the previous phase. Staff read EPEs on related topics when new projects are being designed, but this is not mandatory at KfW and is in fact rarely done according to the interviewees. In addition, training sessions are held by the FC Evaluation Unit at which current findings are reported. Every two years, the FC Evaluation Unit publishes a synthesis report containing the most important overarching findings and brief summaries of interesting EPEs from the past two years. Most people surveyed in operational areas said that they read this synthesis report (KFW2, KFW3, KFW4, KFW5, KFW6).

A lack of time seems to be the main reason why staff do not read reports very much or at all: in Cracknell (2002, p. 189) and in the study by Krämer et al. (2021, p. 42), staff at DC organisations – like the people interviewed for this study – say that they do not have time to read evaluation reports. They need to read a great deal of documents during their daily work. Moreover, the operational staff are concerned with their own projects and do not have time to look at the EPEs of other projects, even if they might be interesting (KFW2, KFW6). Moreover, KFW4 says that no one is interested in the EPE and therefore no one reads it unless there are overarching conclusions or follow-on projects, as staff already have a considerable workload. It would thus appear that few new insights can be gained from the EPEs for the operational areas.

6.2 How does KfW use its ex-post evaluations?

Benefit for staff

According to some of the interviewees, there are various uses for the EPEs that concern KfW staff directly. For example, the EPE reflects KfW staff members' own work in the project (KFW1, KFW2). It ensures that staff take a critical look at their own work (KFW4) and are able to assess their own performance (KFW6) – providing that they read the EPE. However, EPEs have little impact on the design and implementation of current projects (KFW6). KFW5 can draw inspiration from reading the EPEs of innovative or successful projects or seeing presentations on them and hopes that the EPEs lead to an improvement in the quality of KFW5's own projects.

Due to staff rotation at KfW (and at BMZ), EPEs are also used for knowledge management (BMZ1). This is relevant to KFW5 too, as the portfolio managers regularly rotate at KfW and EPEs help them understand the work of their predecessors. Knowledge is transferred (KFW3) in a way that is useful for operational staff.[25] Direct internal effects of poor ratings on individual careers or the reputation of a KfW division were not observed. According to the interviewees, staff rotate too frequently for this to happen.

Using ex-post evaluations to build credibility with BMZ

EPEs can provide in-depth insight and hence facilitate reflection processes that can highlight and enhance the legitimacy of FC work with BMZ as the commissioning party. However, this presupposes that the evaluation is transparent and independent. For the FC Evaluation Unit, the evaluation is a balancing act, because it needs to demonstrate and defend its independence outside the unit. “A reputation of producing deliberately biased reports needs to be avoided at all costs” (KFW4). KFW6 explains that the independence of the FC Evaluation Unit and the system of delegates is extensively discussed outside KfW. KFW6 sees little external recognition of good ratings, because an internal evaluation is not regarded as being completely independent in some cases.

Interviewees who are less familiar with EPEs and with KfW’s evaluation system seem to be more negative about their independence and methodology (in this case BMZ1 and IDOS1) than those who frequently work with KfW’s EPEs. For example, BMZ1 says that the implementing organisation leaves itself open to attack if EPEs are not conducted by an external party, while IDOS1 comments that only external evaluators can be independent. This contrasts with statements by the following KfW staff: Despite sometimes knowing delegates, KFW3 believes that the evaluations were still independent, adding that internal evaluators ask more questions and delve deeper than external evaluators, for example. KFW6 agreed with this statement. KFW4 added that the FC Evaluation Unit was at least as independent as external evaluation consultants because the latter depended on receiving follow-on commissions and reservations might therefore be justified that the ratings they gave were more positive than warranted (KFW4).

Accountability/Transparency

Accountability and transparency refer to KfW’s obligation towards BMZ to justify its actions and decisions. In theory, accountability is linked to sanctions and is based on incentives (Klingebiel, 2013, p. 207). According to the interviewees, however, sanctions are rarely imposed for poor ratings.

In order to guarantee transparency, it needs to be clear how an EPE has been conducted. According to DEV1, a transparent EPE needs to specify a clear question, outlining the data collection methodology and the method of analysis (DEV1). Some information, such as the methodology contained in Part 2, is not transparent for the public, however. The partner or lead executing agency, too, only has access to Part 1, not the full report. Noltze et al. recommend publishing the evaluation reports in full in the interests of transparency (Noltze et al. 2018a, p. 47). Data protection and transparency are conflicting objectives in this context, but in some cases this is merely used as an excuse. It would therefore not be a problem to ensure transparency. If the EPE contains information of a politically highly sensitive nature or data protection is explicitly jeopardised, this (small) area could be redacted in the published version.

The accountability and learnings functions are also conflicting objectives, as already outlined in Section 3. In 2001, Borrmann wrote that EPEs and the associated impact studies in German

25 Further lessons learned in the operational areas are presented in Section 6.5.

DC are essentially understood to be a means of achieving accountability and not as a learning tool. According to Borrmann, EPEs are rarely used to assist in planning new DC projects (Borrmann, 2001, p. 17). In 2009, however, Borrmann and Stockmann described an opposite trend, noting that evaluations were largely used for learning and less commonly for accountability purposes (Borrmann & Stockmann, 2009a, p. 152). In contrast, a meta-evaluation by Kirsch and Wilson on projects in Afghanistan states that "evaluation results serve more on proving accountability with regard to the contributions made and less on drawing insights which encourage critical learning and support portfolio management" (Kirsch & Wilson, 2014, p. viii). The OECD (2021b, pp. 99–100) arrives at a similar conclusion for German DC. This is an observation made outside Germany too: according to Yanguas (2021, pp. 3–4), Forestieri (2020, p. 72) and Reinertsen et al. (2022, p. 366), evaluations in the DC organisations they studied are primarily used for accountability purposes.[26]

With regard to the EPEs conducted by KfW, all interviewees and the KfW website confirm that they are mainly used for learning purposes. If we look more closely, however, we can see an emphasis on accountability from the random sample and from the fact that the EPEs are barely read and – at least at an institutional level – little is learned from them.[27] Nevertheless, the FC Evaluation Unit is attempting a balancing act here in that interesting projects not included in the random samples are also evaluated. As the EPEs try to serve both functions, the fact that there is a clash between the objectives is an additional reason why the EPEs might not be used for learning purposes.

6.3 The political impact of a high success rate

Since 2007, KfW has been calculating the success rate from the overall ratings of all the EPEs conducted (KfW, 2014, p. 48). In order to assess how useful success rates are, one of the aspects that needs to be examined is whether EPEs and the ratings they contain are comparable to enable an aggregated rate to be determined, because – as described in Section 5 – EPEs are partly subjective assessments. Some sectors have a lower average rating than others, e.g. due to a higher risk (KFW6). Here, too, we therefore need to look at how useful it is to group all the sectors together to calculate a success rate. In addition, it should be remembered that success rates are a dimension – often clearly visible – for assessing projects, because ratings are easy to communicate. Yet key learning processes documented in the EPEs can primarily be pursued in individual projects with a *negative* rating. The visibility of these learning processes is limited by the fact that communicating negative assessments is more complex, however.

According to KfW, the size of the samples means that the success rate can be reliably determined and that it is a relevant reflection of all the projects that have been completed (KfW, 2014, p. 48). However, it warns against misinterpretation, explaining that the success rate can only provide limited information about the actual quality of FC, because external factors also need to be taken into account in the success or failure of projects (KfW, 2014, p. 41). Terberger, too, warns that success rates should be interpreted with caution, noting that they cannot be compared between different institutions, as they might be based on different evaluation systems, and that comparing EPEs from different regions and sectors within an institution also offers only limited insight (Terberger, 2011, pp. 233–234).

KFW6 rejects the idea that the ratings are comparable, saying that they should not be compared between the different sectors in particular, but instead only used to find weaknesses and

26 The study by Yanguas (2020) examined DC by the United Kingdom and the World Bank, while Forestieri (2020) looked at Italian DC and Reinertsen et al. (2022) at Swedish and Norwegian DC.

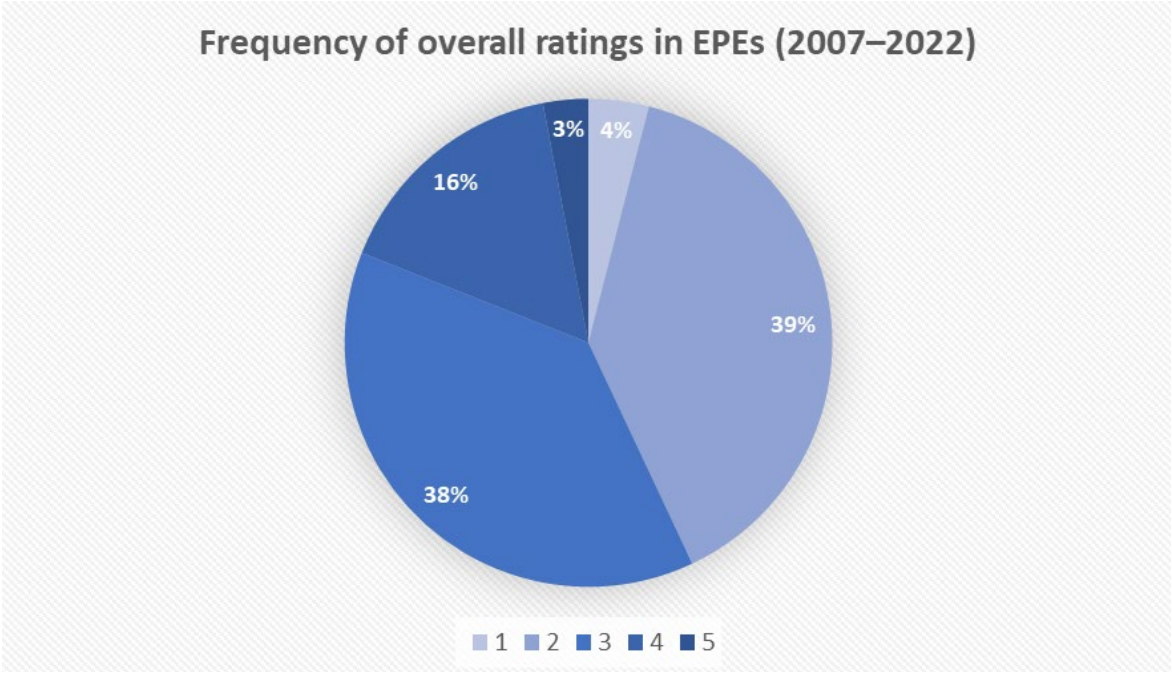
27 This is described in detail in Section 6.5.

improve specific projects (KFW6). KFW6's comments were echoed by other interviewees. Despite addressing the same question and using similar data collection and analysis methods, it can be a challenge, especially for delegates with less experience, to ensure that one's own subjective perspective does not play a role.

With regard to the distribution of ratings awarded by KfW, it is notable that a rating of 2 or 3 is most common, while few projects are given a rating of 1 (KFW6). DEV1 adds that only few projects are given a rating of 1 or 5, and none a 6. DEV1 explains that DC operates in high-risk contexts and thus questions the high success rate. On the other hand, DEV1 believes that there is a path dependency here: as very positive ratings have been awarded over a long period, it would be difficult to downgrade the ratings at the present time (DEV1).

As can be seen in Figure 1, ratings of 2 and 3 are the most common ones. Ratings of 4 or less mean that the project is not regarded as successful. A total of 81% of projects have been rated as successful since 2007, and 19% as unsuccessful. Since it was first determined in 2007, the annual success rate has fluctuated between 77% and 87%, whereby the confidence intervals overlap at around 81% (KfW, 2021a, pp. 42–45). It should be emphasised that a "successful" rating based on the DAC criteria does not necessarily equate with a project's long-term effectiveness. In order to demonstrate the latter, a rigorous impact evaluation would need to be carried out because, as already explained, it is often not possible to make anything more than assumptions on the DAC criteria "development impact" and "sustainability".

Figure 1: Frequency of overall ratings in EPEs from 2007 to 2022 at KfW



Source: Author; data: KfW, 2022d

The average success rates have changed little over the years. Progress is observed at the micro level, but this is scarcely reflected at the macro level (IDOS1). This realisation can be attributed to the micro-macro paradox identified by Mosley (1986), according to which individual development projects have a high success rate but hardly any impact can be seen as a result of DC in the partner countries at the macro level. This is regarded as being due to insufficiently precise questions (in evaluations) concerning impact at the micro level. Caspari and Barbu (2008, p. 2) refer to this as the "evaluation gap", as the methods used in evaluations are not adequate to demonstrate the outcome and impact of development projects precisely. Terberger points out that a good evaluation is not solely concerned with measuring the impact but also with

asking the right questions and avoiding drawing hasty conclusions (Terberger, 2011, p. 237). She adds that calculating the success rate is less important than determining what causes a project to be successful or not (Terberger, 2011, p. 235). Nevertheless, EPEs should particularly assess the effectiveness of projects, because this is ultimately what DC aims to improve.

DEV1 talks about the high success rate as follows:

> In principle, this means that 85% of all FC projects are successful. And you can think that's great. But you also need to ask yourself by way of comparison: If you take a start-up in Silicon Valley – they have excellent conditions to be successful... And DC operates in contexts that couldn't really be worse. … and these very projects have a much higher success rate than the tech start-ups in Silicon Valley. Well... in principle, something's not right there… Because if it [the success rate] were to decrease, everyone is obviously going to ask: "Why are you getting worse?" But I think this levelling… at some point we really need to think about how credible it is that 85% of the projects are considered a success. Whereby a rating of 3, you have to remember, means that not everything is good… And there aren't many 1s. But you also have to ask yourself why there is never a 6 and very rarely a 5. Somehow something's not right with the scale. (DEV1)

The success rate shapes the public's view of DC and – in the case of KfW – of FC (IDOS1). It should be viewed critically, however, because it is much less informative than may appear at first glance. The success rate says nothing about the development effectiveness. It is determined despite a lack of comparability in some cases, even if – according to the people interviewed for this study – it paints slightly too positive a picture and gives the impression that countries and sectors are comparable when in fact they are not. This positive impression is linked to a certain degree of path dependency, because after 15 years of consistently awarding good ratings, it would be difficult to justify to BMZ or the public why they should be modified.

6.4 Relevance to political steering: does BMZ use ex-post evaluations?

The Dexis Consulting Group recommends that DC should open itself up to predominantly evidence-based decision-making. Yet there are difficulties involved in devising effective strategies to ensure that evidence is integrated into policy (Dexis Consulting Group, 2020, p. 39). The study by Krämer et al. also shows that evidence is often not used for strategic decision-making at BMZ (Krämer et al., 2021, p. vi). The authors argue that this is due to German DC being divided up into political steering and implementing organisations. This prevents evidence from the project level being fed into strategic decision-making (Krämer et al., 2021, p. 42). BRH criticises that reporting at project level makes it almost impossible to determine relevance and impact at programme level. BMZ rarely carries out an overarching analysis of findings from individual project evaluations (BRH, 2021, p. 7). It therefore has few options for steering policy based on evidence at programme level (BRH, 2021, p. 24).

BMZ says itself that it aims to improve political and strategic steering by using aggregated findings from project evaluations (BMZ, 2021a, p. 40). According to the interviews conducted for this study, project evaluations focus too much on small-scale details for BMZ to be able to use them to design programmes or country portfolios. Yet even aggregated evidence only appears to be harnessed for policy-making to a limited extent and is perhaps seen as less of a priority at BMZ. Evaluations there are thus regarded as administrative tasks used for formal accountability, but not as a learning tool (Krämer et al., 2021, p. 41; cf. OECD, 2021b). Even the OECD (2001, p. 72) concludes that the findings obtained from evaluations are only a small component that is fed into political decision-making. Despite learning and feedback loops, however systematic they may be, a range of other factors – such as (political) power struggles, unfavourable political situations in the partner country, spending pressure or decision-makers'

own interests – have a greater influence on the political steering of DC programmes (OECD, 2001, p. 72).

Failure to coordinate time schedules between project cycles and evaluations also makes it more difficult to take up findings at BMZ (Krämer et al., 2021, p. 41), as these findings – like those of EPEs – are not available until (long) after the project has been completed. BMZ1 confirms that the findings often arrive at the wrong time for BMZ when no current negotiations are about to take place with the partner concerned. The EPEs are therefore filed away or only skimmed over. According to BMZ1, they are designed (due to the time lapse between the end of the project and the evaluation) to reinforce the impression that already prevailed when the project was completed and that the other reports on the project convey. BMZ1 therefore does not expect to obtain any major new insights from EPEs.

With regard to the process, the evaluation reports (like all other reports) are “acknowledged” by the country division. Taken together, these reports can have consequences for the future design of the portfolio, for example whether there are more activities in this area or certain things need to be adapted (BMZ1). BMZ1 acknowledges that EPEs are important when the next negotiations are held between BMZ and the partner, but is not really interested in reviewing the past but in looking at what could be done better in the future.

It becomes evident that different evaluation findings are relevant at different levels, e.g. for political steering or making technical improvements to implementation. EPEs appear to have only limited influence on political steering. It is noteworthy that according to BMZ1, EPEs are primarily relevant to the implementing organisations, whereas several operational staff members at KfW believe that EPEs are mainly relevant to BMZ.

6.5 So who is learning from ex-post evaluations?

KfW particularly emphasises the learning function of EPEs, as outlined in Section 3. Yanguas distinguishes between operational and strategic learning. Operational learning is about the process of project implementation. Strategic learning allows development organisations to realign their processes, structures and practices to ensure a more effective pursuit of organisational goals (Yanguas, 2021, pp. 4–5). The two types of learning do not necessarily feed into or reinforce one other: micro-learning from operations rarely leads to a macro-shift in policy (Yanguas, 2021, p. 5). Strategic learning can be equated with institutional learning, whereas operational learning is described in the following as “learning by operational areas”. At this point, the level of individual/personal learning by KfW staff through the delegates system should also be distinguished.

Institutional learning at KfW

KfW states that institutional lessons learned are to be identified for future projects (KfW, 2021a, p. 1). All projects in the random sample are evaluated, even if some of the projects are ones in which there are no more lessons to be learned because similar project designs have already been evaluated or this type of project is no longer being implemented (KFW4).

According to KfW (2021a, p. 1), institutional learning takes place by the findings of the evaluation being processed in a way that is appropriate for the target groups and can be used efficiently. Even if projects do not have a follow-on phase, the option of transferring lessons learned to similar contexts should be created (DEV1). The FC Evaluation Unit produces cross-sectional analyses – stratified by sector or country – supported in some cases by consultants or universities. KFW3 explains that there is some degree of knowledge transfer: experience acquired in projects implemented in Asia is now being applied in Africa, for instance. Although some interviewees mention examples of institutional lessons learned from EPEs, they still say

that this rarely happens, partly because of the problems involved in transferring them to other contexts.

Learning by operational areas

The interviews in the study by Krämer et al. on the topic of rigorous impact evaluations indicate that insights gained from them are barely used for operational learning, even though this was the reason most commonly cited for using rigorous findings (Krämer et al., 2021, pp. 35–36). A similar observation regarding EPEs was also made in the research for this paper: the interviews conducted for this study suggest that EPEs are not particularly important for operational work. Most aspects are not new for the portfolio managers and technical experts (KFW4). They are already aware of the recommendations and criticisms (KFW6) and therefore learn little from EPEs. Stockmann confirms the argument that evaluation findings contain nothing new for operational areas (Stockmann, 2007, p. 180).

Yanguas (2021, p. 15) answers the question as to why little operational learning takes place using four arguments. Firstly, the tools and methods used are not easy to replicate in different contexts. This argument is also cited by the interviewees: KFW1 assumes that other evaluations may contain points that can be used in one's own project. However, KFW1 thinks that it would be complicated to transfer findings to other contexts. Although some mistakes are repeated, they have a different impact in different contexts. Not all mistakes can be avoided, as projects work with a partner that has or ought to have a considerable degree of autonomy. KfW, too, states in its biennial synthesis report that the statements and recommendations made in evaluations cannot be readily transferred to other projects (KfW, 2019, p. 61).

Secondly, the implementing organisations have an incentive to hoard knowledge as they compete with one another for contracts and grants. This argument does not really apply in German DC: technical and financial cooperation differ in the types of projects they implement and the expertise required, so KfW does not compete for funding. At KfW, knowledge hoarding can be attributed to a lack of transparency in the EPEs. Thirdly, there is an incentive not to publish complications that occur in projects in the partner countries (and therefore their solutions) to avoid drawing public attention to them. The final argument put forward by Yanguas is that all reporting – including evaluations – is not fed back to practitioners in a systematic, operationally relevant manner (Yanguas, 2021, p. 15). Although this does happen at KfW, the reports are seldom read. Ramalingam (2005, p. 30) adds that learning is one of several tasks competing for staff members' attention. This problem is also reflected in the interviews with operational staff at KfW.

If an evaluated project has follow-on phases, lessons can be learned from the EPE of the earlier phases (e.g. in terms of indicators and objectives). KFW1 and KFW2 reported that the EPE influenced the indicators and the results matrix of their current follow-on project, for example (KFW1, KFW2). KFW4 also confirmed that an evaluation may lead to adjustments being made in an ongoing project. If there are follow-on projects or further cooperation in the sector, partners and portfolio managers are both more interested in the EPE (KFW4). If cooperation is being phased out, little interest is shown and there are only limited lessons learned (KFW4). According to KFW2, the EPE is of no interest if there are no follow-on projects.

When portfolio managers plan new projects at KfW, they can now use a recently introduced software program called Quick Evaluation Results, in which all EPEs are stored and sorted by country, sector and lessons learned. However, there is no obligation to consult old EPEs when designing new projects. Most of the interviewees in operational areas do not commonly do so either: KFW2 and KFW6 say that past EPEs often have little influence on the design or implementation of current projects because they cannot obtain any more new insights from them. KFW5 was the only person who said that they actively searched for previous EPEs by

lead executing agency, sector and country and obtained information from them when preparing projects.

In order to disseminate lessons learned within KfW, staff from the FC Evaluation Unit take part in sector seminars in operational areas and present the current evaluation findings there (KFW4, KFW5), for example with regard to designing and planning future projects (KFW3). There is no such format at KfW for BMZ or for the partners and lead executing agencies. The biennial synthesis report is also used for learning purposes. The FC Evaluation Unit disaggregates and publishes the EPE findings by country and sector (KFW2).

Learning by the delegates

An analysis of the interview data shows that EPEs are primarily used for learning through the system of delegates. The idea of this system is to feed experience from conducting an EPE into the everyday operational work of delegates and to raise their awareness of the topic of impact and of measuring and documenting evidence of results (KfW, 2021a, p. 3). The individuals interviewed for this paper confirmed this learning effect among delegates in the course of their work. Delegates increase their knowledge about sectors and countries, and the process of conducting the EPE serves as internal training (KFW4). Acquiring knowledge in connection with the results matrix in particular is very useful for portfolio managers in their daily work, for example, as they can use this knowledge when designing new projects. The interviews with KfW by Borrmann and Stockmann also emphasise the learning aspect among delegates as being particularly interesting, which is in line with the considerable interest on the part of staff acting as delegates (Borrmann & Stockmann, 2009b, p. 261). BRH also acknowledges that the delegates system is useful with regard to the lessons learned by delegates (BRH, 2022, p. 10).

“Learning by doing” is the most effective way to learn (OECD, 2001, p. 27). It therefore makes sense to involve the main stakeholder groups in the evaluation process, as they have the greatest exposure to the lessons being learned, can internalise these lessons and feed them into the next stage of planning (OECD, 2001, p. 27). It is not only the report that facilitates learning; the process of producing an evaluation also promotes learning among those involved (Borrmann & Stockmann, 2009a, p. 133). Nevertheless, it should be noted that delegates have the greatest potential to distort the findings of an EPE as a result of their methodological deficits (see BRH, 2022, p. 11). Here, too, there are conflicting objectives between the individual actors who learn from the EPE (the delegates) and the credibility and independence of the evaluation.

In summary, there are various reasons why EPEs result in little or no learning. Yanguas (2021) explains that the knowledge thus generated is not or only rarely used. The information compiled in the EPE can be used for learning, but the cycle is frequently interrupted and ends with the evaluation because people do not read the report and so the evaluations do not lead to learning (IDOS1). No incentives are provided for learning in development organisations: “Part of the challenge lies in convincing staff that knowledge sharing means smarter work rather than more work” (King & McGrath, 2003, p. 15).

The main reason why the EPEs result in only a limited learning effect among the projects is the considerable time lapse. KFW4 states that the operational staff already have access to more advanced tools at the time of the EPE: “At least that was the feedback we were often given during my time in the FC Evaluation Unit: We’d never use something like that nowadays, because we’ve already learned from our mistakes” (KFW4). KfW is aware that the time lapse creates obstacles: no short-term findings can be obtained for designing future projects (KfW, 2021c, p. 4). If EPEs were to be conducted sooner, however, they would not be able to provide any information about sustainability or development impact.

Interviews and literature ultimately suggest that there is little operational or institutional learning from EPEs, especially not from reading the report. Delegates appear to be the only people to

learn from EPEs – the ones that they conduct themselves. EPEs may lead to improvements in follow-on phases or projects, if there are any. It does not make sense to address the problem that the findings are published much later, for example, because of the competing objective of adequately evaluating the sustainability and development impact of a project. It would therefore appear to be a better idea to continue using EPEs to achieve accountability and as raw material for further research and meta-evaluations.

7 How ex-post evaluations can be used in future

Overall, this study shows that despite all the challenges involved, EPEs are an important tool for accountability and can help enhance the legitimacy of FC. It should be emphasised that EPEs appear to be conducted with care at KfW and are taken seriously by the staff interviewed here. All the people interviewed during the research for this paper were of the opinion that EPEs are *basically important*.

However, the methodology of the EPEs creates challenges. They are largely connected with the person conducting the evaluation and how they deal with it. If evaluators have less experience of applying evaluation techniques or lack prior knowledge of sectors and countries, this may bias evaluation findings. The subjectivity of evaluators is another issue: delegates can evaluate individual projects independently, but as members of the KfW workforce they are part of the overall system of project design and implementation. This system could be critically examined more reliably and with greater impartiality by external evaluators. For this reason, the comparability of the individual EPEs can pose a challenge.

Transparency and data protection are conflicting objectives: the importance that KfW places on data protection in EPEs limits their transparency and hence credibility in the eyes of the public. At the same time, data protection guarantees safety for the partner, who might otherwise be open to attack on the basis of the information contained in the EPE. Nevertheless, taxpayers' money needs to be managed transparently, which is why a transparent evaluation of projects is recommended, as could be done through EPEs.

Even though this study is not representative, the statements made by the interviewees indicate that EPEs are only rarely read by operational staff at KfW, so they can have little impact there: reading the reports ought to be a basic requirement for all subsequent learning processes. However, merely reading a report does not automatically lead to learning, if the insight drawn from it is not applied. If operational staff regarded the content of EPEs as being relevant to their work, they would presumably read the reports in more detail. In fact, according to the interviewees, EPEs are regarded by operational staff as largely irrelevant to their own work.

Around 19% of the evaluated projects are given a rating of 4 or worse, in other words they are regarded as "not successful". As far as could be established here, projects that are "not successful" have little impact within KfW with regard to personal careers, the reputation of a KfW division, a reduction in funding or termination of cooperation with a partner. If poor evaluation ratings have no consequences, there is no incentive to manipulate ratings in order to improve them. This supports the accountability function of the EPEs. On the other hand, however, this means that there is little incentive to tackle criticisms, so the learning function is barely addressed.

The two most important functions of EPEs are accountability and learning, which, as shown above, are conflicting objectives. They cannot both be achieved at the same time using the same methods: a random sample guarantees that results are representative, for example, thus contributing to accountability. As a result, however, evaluations are often of project types that have already been evaluated many times or whose approaches are not innovative. According to interviewees, this means that lessons are less frequently learned. KfW's delegates system is

also connected to this, as internal evaluators should be used, where possible, in the interests of learning to ensure that lessons learned can be put into practice more effectively. In contrast, external and hence more independent evaluators are more suitable for the accountability function (cf. also Reinertsen et al., 2022, p. 364).

According to the interviewees, the time lapse means that the EPEs do not foster learning in operational areas. Without the time lapse, however, sustainability and development impact could not be adequately evaluated. The greatest learning effect appears to be seen among delegates while they are carrying out an EPE. However, they may cause biases in the EPE. BRH therefore recommends additional training sessions for delegates (BRH, 2022, p. 11). Commissioning external evaluation consultants does not appear to be a better solution, as they, too, might have an incentive to give more positive ratings.

Staff at KfW and BMZ cannot learn from EPEs if they do not read them or the findings are not otherwise registered (e.g. in accumulated form). In recent years, the FC Evaluation Unit has tried to introduce innovative formats to address this problem within KfW. The dissemination of evaluation findings is still too seldomly seen as a further important evaluation phase, however. If there are follow-on phases or projects, the interviewees from KfW's operational areas are most likely to benefit from the EPEs, as the criticisms can be applied to the project immediately. Particularly if cross-sectional evaluations are generated from the EPEs, these may lead to overarching lessons learned.

This analysis suggests that EPEs have only limited potential to improve future FC projects. Instead, their use is restricted mainly to accountability. At BMZ, too, the benefit and use of EPEs would appear to be fairly limited. Nevertheless, getting rid of EPEs altogether does not appear to be a solution either, as they are the only instrument that can be used to evaluate sustainability and development impact of a representative number of projects relatively cost-effectively at KfW, hence providing a basis for achieving accountability in FC. EPEs fulfil only a small part of the function originally intended by BMZ. The question thus arises as to whether the purposes that an EPE is designed to fulfil should be reconsidered.

As some of the conflicting objectives presumably cannot be reconciled, the problem of pursuing many different goals remains. There is feasible potential for improvement in increasing the learning effect from EPEs, however: the author of the present paper agrees with BRH and the OECD, which recommend (to KfW) that overarching conclusions from the EPEs should be compiled more often in cross-sectional analyses (BRH, 2022, p. 6; OECD, 2001, pp. 27–28) with a view to improving institutional learning at KfW. These analyses in particular should be addressed in the biennial synthesis report. According to some of the interviewees, operational staff learn in other ways (e.g. from the synthesis report mentioned above or from information events) and not from individual EPEs.

In order to continue developing a learning system, it is also recommended that the methodology section of each EPE should be made available to the public – or even Part 2 in full, as this would mean that other research institutes or organisations could access it without any bureaucratic hurdles and could produce meta-evaluations or cross-sectional analyses independently. Secondly, it would seem to make sense for the country and sector divisions at BMZ to produce cross-sectional analyses from EPEs, too, as there is currently no standardised method for using EPE findings at this level (BRH, 2021, p. 20). To do so, Krämer et al. propose setting up additional platforms for sharing ideas between political decision-makers, development practitioners, researchers and evaluators to ensure that evidence is generated in line with actual needs (Krämer et al., 2021, p. 59). Another proposal involves including KfW's partners and lead executing agencies in this exchange, thus allowing them to learn from the EPEs or other evaluation formats too. In addition, a support structure for practitioners and decision-makers should be set up covering the entire cycle of evidence generation and use; the structure would be partly decentralised and partly based at a central level in order to be able to work close to

the practical level while at the same time offering cross-organisational learning and advisory services (Krämer et al., 2021, p. 60). This support structure could also be in line with the proposal by BMZ1, who would like there to be presentations of the findings of EPEs (or of overarching conclusions) to facilitate a better understanding of these findings in BMZ's country and sector divisions.

References

Amine, M., & Eulenburg, A. G. zu (2022). *Wirkungsorientierte bilaterale Entwicklungszusammenarbeit? Der Dreiklang aus Länderstrategie, EZ-Programm und Modul* (DEval Policy Brief 7). Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval).

Armytage, L. (2011). Evaluating aid: An adolescent domain of practice. *Evaluation*, *17*(3), 261–276. https://doi.org/10.1177/1356389011410518

Bartl, W., Papilloud, C., & Terracher-Lipinki, A. (2019). Governing by numbers – key indicators and the politics of expectations: An introduction. *Historical Social Research*, *44*(2), 7–43. https://doi.org/10.12759/HSR.44.2019.2.7-43

Beebe, J. (1995). Basic Concepts and Techniques of Rapid Appraisal. *Human Organization*, *54*(1), 42–51. https://doi.org/10.17730/humo.54.1.k84tv883mr2756l3

BMZ (Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung). (2014). *BMZ-Stellungnahme zum DEval-Bericht: „Ein Review der Evaluierungsarbeit zur deutschen Entwicklungszusammenarbeit in Afghanistan“.*

BMZ. (2020). *Evaluierungskriterien für die deutsche bilaterale Entwicklungszusammenarbeit. BMZ-Orientierungslinie zum Umgang mit den OECD-DAC-Evaluierungskriterien in Evaluierungen der deutschen bilateralen Entwicklungszusammenarbeit*.

BMZ. (2021a). *Evaluierung der Entwicklungszusammenarbeit. Leitlinien des BMZ* (BMZ-Strategien 4).

BMZ. (2021b). *Glossar der Schlüsselbegriffe im Bereich Evaluierung der Entwicklungszusammenarbeit*.

BMZ. (2021c). *Leitlinien für die bilaterale Finanzielle und Technische Zusammenarbeit mit Kooperationspartnern der deutschen Entwicklungszusammenarbeit.*

BMZ. (2022). *Deutsche ODA*. https://www.bmz.de/de/ministerium/zahlen-fakten/oda-zahlen/deutsche-oda-leistungen-19220

Borrmann, A. (2001). *Reform der Erfolgskontrolle in der deutschen Entwicklungszusammenarbeit. Eine Zwischenbilanz: Studie im Auftrag des Bundesministeriums für wirtschaftliche Zusammenarbeit und Entwicklung* (HWWA-Studien, Vol. 63, 1. ed.). Baden-Baden: Nomos-Verl.-Ges. http://www.hwwa.de

Borrmann, A., & Stockmann, R. (2009a). *Evaluation in der deutschen Entwicklungszusammenarbeit. Band 1 – Systemanalyse* (Sozialwissenschaftliche Evaluationsforschung). Münster: Waxmann.

Borrmann, A., & Stockmann, R. (2009b). *Evaluation in der deutschen Entwicklungszusammenarbeit. Band 2 – Fallstudien* (Sozialwissenschaftliche Evaluationsforschung, 8,1). Münster: Waxmann. http://www.socialnet.de/rezensionen/isbn.php?isbn=978-3-8309-2125-7

BRH (Bundesrechnungshof). (2021). *Abschließende Mitteilung an das Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung über die Prüfung „Evaluierung von Maßnahmen der Entwicklungszusammenarbeit. Teil I – Steuerung im BMZ“*.

BRH. (2022). *Abschließende Mitteilung an das Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung über die Prüfung „Evaluierung von Maßnahmen der Entwicklungszusammenarbeit. Teil III – Maßnahmen der bilateralen Finanziellen Zusammenarbeit“*.

Caspari, A., & Barbu, R. (2008). *Wirkungsevaluierungen: Zum Stand der internationalen Diskussion und dessen Relevanz für die Evaluierung der deutschen Entwicklungszusammenarbeit* (Evaluation Working Papers). Bonn: BMZ Evaluation Division.

Chambers, R. (1981). Rapid rural appraisal: rationale and repertoire. *Public Administration and Development*, *1*, 95–106.

Cracknell, B. E. (2002). *Evaluating development aid. Issues, problems and solutions* (3rd printing). London: Sage.

Czarniawska, B. (2007). *Shadowing. And other techniques for doing fieldwork in modern societies*. Koege: Copenhagen Business School Press – Universitetsforlage.

Desrosières, A. (2015). Retroaction: How indicators feed back. In R. Rottenburg, S. E. Merry, S. J. Park & J. Mugler (Eds.), *The world of indicators. The making of governmental knowledge through quantification* (Cambridge studies in law and society, pp. 329–353). Cambridge: Cambridge University Press.

Dexis Consulting Group. (2020). *Evidence base for collaborating, learning and adapting. Summary of the literature review*. USAID.

Faust, J. (2020). Rigorose Wirkungsevaluierung: Genese, Debatte und Nutzung in der Entwicklungszusammenarbeit. *der moderne staat: Zeitschrift für Public Policy, Recht und Management*, *13*(1), 61–80. https://doi.org/10.3224/dms.v13i1.08

Ferguson, J., Mchombu, K., & Cummings, S. (2008). *Management of knowledge for development: Meta-review and scoping study* (IKM Working Paper, 1).

Forestieri, M. (2020). Equity Implications Evaluating Development Aid: The Italian Case. *Journal of MultiDisciplinary Evaluation*, *16*(34), 65–90.

GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit). (2012). Wirkungsanalyse und Wirkungsmessung in Gesundheitsvorhaben der deutschen Entwicklungszusammenarbeit. *Zukunftsentwickler. Wir machen Zukunft. Machen Sie mit.*

Glaser, B. G., & Strauss, A. L. (2010). *Grounded theory. Strategien qualitativer Forschung* (Programmbereich Gesundheit, 3. ed.). Bern: Huber.

Goodwin, M. (2017). The poverty of numbers: reflections on the legitimacy of global development indicators. *International Journal of Law and Context*, *13*(4), 485–497. https://doi.org/10.1017/S1744552317000404

Hartmann, C., Amine, M., Klier, S., & Vorwerk, K. (Eds.) (2019). *Länderportfolioreviews. Ein Analyseinstrument für die deutsche Entwicklungszusammenarbeit*. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit. https://nbn-resolving.org/urn:nbn:de:101:1-201909191303236939922 9

Henry, G. T., & Mark, M. M. (2003). Beyond Use: Understanding Evaluation's Influence on Attitudes and Actions. *American Journal of Evaluation*, *24*(3), 293–314. https://doi.org/10.1177/109821400302400302

Hoey, L. (2015). Show me the numbers: Examining the dynamics between evaluation and government performance in developing countries. *World Development*, *70*, 1–12. doi:10.1016/j.worlddev.2014.12.019

Holzapfel, S., & Römling, C. (2020). *Monitoring in German bilateral development cooperation. A case study of agricultural, rural development and food security projects* (Discussion Paper 18/2020). Bonn: German Development Institute / Deutsches Institut für Entwicklungspolitik (DIE).

Hovland, I. (2003). *Knowledge management and organisational learning: An international development perspective* (Working Paper 224). https://www.files.ethz.ch/isn/95711/wp224.pdf

Kalton, G., & Anderson, D. W. (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society*, *Ser. A, 149*, 65–82.

KfW (Kreditanstalt für Wiederaufbau). (2014). *13. Evaluierungsbericht 2013–2014. Wirkung im ländlichen Raum: Tragfähige Ansätze für Mensch und Natur*.

KfW. (2019). *15. Evaluierungsbericht 2017–2018. Zu größerer Wirkung in kleineren Städten*.

KfW. (2021a). *16. Evaluierungsbericht 2019–2020. Begleiten. Bewerten. Lernen.*

KfW. (2021b). *KfW Entwicklungsbank: Zahlen und Fakten 2020*.

KfW. (2021c). *Das Wirkungsmanagement der KfW-Bankengruppe. Wirkungsverständnis, -kategorien, -indikatoren und Zusammenspiel der Daten*.

KfW. (2022a). *Rolle/Wichtigkeit von RIE*. kfw-entwicklungsbank.de/s/dezBLLbP

KfW. (2022b). *Unsere Arbeitsweise*. kfw-entwicklungsbank.de/s/dezRK8Y

KfW. (2022c). *Unsere Wirkungen*. kfw-entwicklungsbank.de/s/dezqUw9

KfW. (2022d). *Weltweites Engagement*. https://www.kfw.de/microsites/Microsite/transparenz.kfw.de/#/start

King, K., & McGrath, S. (2003). *Knowledge sharing in development agencies: Lessons from four cases*. Washington, DC: World Bank.

Kirsch, R., & Wilson, M. B. (2014). *Ein Review der Evaluierungsarbeit zur deutschen Entwicklungszusammenarbeit in Afghanistan*. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval). https://nbn-resolving.org/urn:nbn:de:0168-ssoar-55213-3

Klingebiel, S. (2013). Ergebnisbasierte Ansätze in der Entwicklungszusammenarbeit: Möglichkeiten und Grenzen neuer Ansätze. In R. Öhlschläger & H. Sangmeister (Eds.), *Von der Entwicklungshilfe zur internationalen Zusammenarbeit* (pp. 201–210). Nomos Verlagsgesellschaft mbH & Co. KG. https://doi.org/10.5771/9783845245676-201

Kogen, L. (2018). What have we learned here? Questioning accountability in aid policy and practice. *Evaluation*, *24*, 98–112. doi:10.1177%2F1356389017750195

Krämer, M., Jechel, L., Kretschmer, T., & Schneider, E. (2021). *Rigorous impact evaluation: Evidence generation and take-up in German development cooperation*. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval).

Krohwinkel-Karlsson, A. (2007). Knowledge and learning in aid organisations: A literature review with suggestions for further studies. *SADEV Working Paper*, *1*.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (Grundlagen Psychologie, 6. ed.). Weinheim: Beltz.

Mayring, P. (2010). Qualitative Inhaltsanalyse. In G. Mey & K. Mruck (Eds.), *Handbuch qualitative Forschung in der Psychologie* (1. ed., 2010, pp. 601–613). Wiesbaden: VS Verlag. https://doi.org/10.1007/978-3-531-92052-8_42

Merton, R. K. (1949). Patterns of influence: a study of interpersonal influence and of communications behavior in a local community. In P. F. Lazarsfeld & F. N. Stanton (Eds.), *Communications research 1948-49* (pp. 180–219). New York: Duell, Sloan & Pearce.

Meuser, M., & Nagel, U. (1991). ExpertInneninterviews – vielfach erprobt, wenig bedacht. Ein Beitrag zur qualitativen Methodendiskussion. In D. Garz & K. Kraimer (Eds.), *Qualitativ-empirische Sozialforschung. Konzepte, Methoden, Analysen* (pp. 441–471). Wiesbaden: Springer Fachmedien.

Meyer, W., Bär, T., Faust, J., Jan, S. von, Silvestrini, S., & Wein, S. (2019). Die DeGEval-Standards in der deutschen bilateralen Entwicklungszusammenarbeit. In J. U. Hense, W. Böttcher, M. Kalman & W. Meyer (Eds.), *Evaluation: Standards in unterschiedlichen Handlungsfeldern. Einheitliche Qualitätsansprüche trotz heterogener Praxis?* (1. ed, pp. 165–181). Münster: Waxmann.

Mosley, P. (1986). Aid-effectiveness: The micro-macro paradox. *IDS Bulletin*, *17*(2), 22–27. https://doi.org/10.1111/j.1759-5436.1986.mp17002004.x

Noltze, M., Euler, M., & Verspohl, I. (2018a). *Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit. Meta-Evaluierung*. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval).

Noltze, M., Euler, M., & Verspohl, I. (2018b). *Nachhaltigkeit in der deutschen Entwicklungszusammenarbeit. Meta-Evaluierung, Zusammenfassung*. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval).

OECD (Organization for Economic Co-operation and Development). (1991). *Principles for evaluation of development assistance*. Paris: OECD Publishing. https://www.oecd.org/development/evaluation/2755284.pdf

OECD. (2001). *Evaluation feedback for effective learning and accountability* (Evaluation and aid effectiveness, vol. 5). Paris: OECD Publishing.

OECD. (2010). *Qualitätsstandards für die Entwicklungsevaluierung* (DAC-Reihe Leitlinien und Grundsatztexte). Paris: OECD Publishing. https://doi.org/10.1787/9789264085183-de

OECD. (2020). *Better criteria for better evaluation: Revised and updated evaluation criteria*. Paris: OECD Publishing.

OECD. (2021a). *Applying evaluation criteria thoughtfully*. Paris: OECD Publishing. https://doi.org/10.1787/543e84ed-en

OECD. (2021b). *DAC-Prüfbericht über die Entwicklungszusammenarbeit: Deutschland 2021. Wichtigste Ergebnisse und Empfehlungen*. Paris: OECD Publishing. https://doi.org/10.1787/83f90077-de

OECD/DAC Network on Development Evaluation. (2019). *Better criteria for better evaluation: Revised evaluation criteria definitions and principles for use*. Paris: OECD Publishing. https://www.oecd.org/dac/evaluation/revised-evaluation-criteria-dec-2019.pdf

Porter, T. M. (2015). The Flight of the Indicator. In R. Rottenburg, S. E. Merry, S. J. Park & J. Mugler (Eds.), *The world of indicators: The making of governmental knowledge through quantification* (Cambridge studies in law and society, pp. 34–55). Cambridge: Cambridge University Press.

Ramalingam, B. (2005). *Implementing knowledge strategies: Lessons from international development agencies* (RAPID Working Paper 244, Vol. 244). London.

Reinertsen, H., Bjørkdahl, K., & McNeill, D. (2022). Accountability versus learning in aid evaluation: A practice-oriented exploration of persistent dilemmas. *Evaluation*, *28*(3), 356–378. https://doi.org/10.1177/13563890221100848

Rottenburg, R., Merry, S. E., Park, S. J., & Mugler, J. (Eds.) (2015). *The world of indicators: The making of governmental knowledge through quantification* (Cambridge studies in law and society). Cambridge: Cambridge University Press.

Sabbi, M., & Stroh, A. (2020). The “Numbers Game”: Strategic reactions to results-based development assistance in Ghana. *Studies in Comparative International Development*, *55*(1), 77–98. https://doi.org/10.1007/s12116-019-09296-z

Schönhuth, M., & Jerrentrup, M. T. (2019). *Partizipation und nachhaltige Entwicklung: Ein Überblick*. Wiesbaden: Springer Fachmedien.

Shallwani, S., & Dossa, S. (2023). Evaluation and the White Gaze in International Development. In T. Khan, K. Dickson & M. Sondarjee (Eds.), White saviorism in international development. Theories, practices and lived experiences (pp. 42–62). Wakefield, Quebec, Kanada: Daraja Press.

Stockmann, R. (2004). *Was ist eine gute Evaluation? Einführung zu Funktionen und Methoden von Evaluationsverfahren* (CEval-Arbeitspapiere 9). Saarbrücken: Centrum für Evaluation.

Stockmann, R. (2007). *Handbuch zur Evaluation. Eine praktische Handlungsanleitung* (Sozialwissenschaftliche Evaluationsforschung, Vol. 6). Münster, München, Berlin: Waxmann. von http://www.socialnet.de/rezensionen/isbn.php?isbn=978-3-8309-1766-3

Terberger, E. (2011). Evaluierung in der Entwicklungszusammenarbeit. Das Beispiel der Finanziellen Zusammenarbeit. Theoretische Konzepte, strukturelle Herausforderungen und praktische Umsetzung. In J. König & J. Thema (Eds.), *Nachhaltigkeit in der Entwicklungszusammenarbeit. Theoretische Konzepte, strukturelle Herausforderungen und praktische Umsetzung* (Globale Gesellschaft und internationale Beziehungen, 1. ed., pp. 219–238). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93091-6

Vindrola-Padros, C., & Johnson, G. A. (2020). Rapid Techniques in Qualitative Research: A Critical Review of the Literature. *Qualitative Health Research*, *30*(10), 1596–1604. https://doi.org/10.1177/1049732320921835

Weber, M. (1977). Objectivity of Social Science and Social Policy. In F. R. Dallmayr & T. A. McCarthy (Eds.), *Understanding and Social Inquiry* (pp. 24–37). Notre Dame: University of Notre Dame Press.

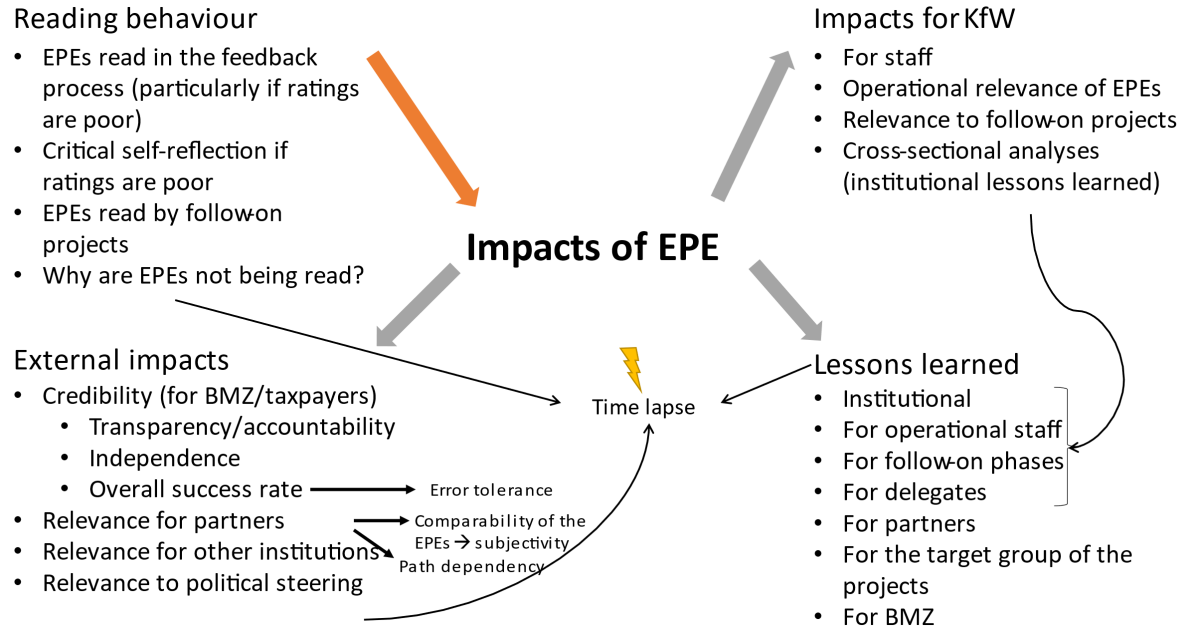
Wilhelm, J. L. (Ed.). (2015). *Evaluation komplexer Systeme. Systemische Evaluationsansätze in der deutschen Entwicklungszusammenarbeit* (Potsdamer Geographische Praxis, Vol. 10). Potsdam: Univ.-Verl.

Yanguas, P. (2021). *What have we learned about learning? Unpacking the relationship between knowledge and organisational change in development agencies* (Discussion Paper 9/2021). Bonn: German Development Institute / Deutsches Institut für Entwicklungspolitik (DIE). https://doi.org/10.23661/DP9.2021

Zintl, M. (2009). Evaluierung in der deutschen Entwicklungszusammenarbeit. In T. Widmer, W. Beywl & C. Fabian (Eds.), *Evaluation.* VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91468-8_24

Annexes

Annex 1: Brief coding system



Source: Author

Annex 2: Information about the interviewees

Interviewee	Background	Has produced an EPE at KfW
KFW1	KfW – operational area	no
KFW2	KfW – previously operational area; now competence centre	yes
KFW3	KfW – operational area	no
KFW4	KfW – previously FC Evaluation Unit; now operational area	yes
KFW5	KfW – operational area	no
KFW6	KfW – operational area	no
BMZ1	BMZ – country officer	-
DEV1	DEval – Competence Centre for Evaluation Methodology	-
IDOS1	IDOS – research on development cooperation	-