I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16691

Volume, Risk, Complexity: What Makes Development Finance Projects Succeed or Fail?

Yota Eilers
Jochen Kluve
Jörg Langbein
Lennart Reiners

DECEMBER 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16691

Volume, Risk, Complexity: What Makes Development Finance Projects Succeed or Fail?

Yota Eilers
University of Oxford

Jörg Langbein
KfW Development Bank

Jochen Kluve
Humboldt-Universität zu Berlin, KfW Development Bank and IZA

Lennart Reiners
Asian Development Bank

DECEMBER 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Volume, Risk, Complexity: What Makes Development Finance Projects Succeed or Fail?*

In 2022, governments around the world committed USD 211 bn. to official development assistance. Despite these high contributions, systematic assessments of the determinants of success—or failure—of development aid projects remain limited, particularly for bilateral development aid. This paper provides such a systematic, quantitative analysis: we construct a unique database covering 5,608 evaluation results—success ratings—for bilateral development aid projects financed through one of the biggest global donors, KfW Development Bank. Detailed data on project characteristics allow us to link success ratings to five clusters of key explanatory factors along the entire project life-cyle and context: (a) In terms of project financing, we find a statistically significant positive association between the financial budget volume of the project and its success ratings, ceteris paribus. Second, concerning the (b) project structure, the type of project partner—government, private sector, multilateral organizations—shows no significant association with project success, suggesting that project implementation works equally well with different partners. (c) Project complexity as measured by both technical complexity and longer implementation duration exerts a negative influence on success ratings. Regarding (d) project risks, a highly relevant and significant predictor for less successful projects is the share of ex-ante identified risks that eventually materialized—suggesting that project designs correctly identify the relevant risks in advance, but are not able to mitigate (all of) them during execution. Finally, concerning (e) the project context there is some indication that higher GDP growth rates are positively associated with project success.

JEL Classification: C40, F35, O10, O19

Keywords: development finance, OECD DAC evaluation criteria, meta analysis

Corresponding author:
Jochen Kluve
Humboldt-Universität zu Berlin
School of Business and Economics
Spandauer Str. 1
10178 Berlin
Germany

E-mail: jochen.kluve@hu-berlin.de

* We gratefully acknowledge valuable comments by David Card, Andreas Fuchs, Krisztina Kis-Katos, Bernhard Reinsberg, seminar participants at University of Göttingen, University of Toronto, the Digital Development Dialogue (3D) and KfW Development Bank, as well as conference participants of the Poverty Reduction, Equity and Growth Network 2022 Kampala, the European Public Choice Society 2023, and the German Development Economics Conference 2023 Dresden. We thank Marlon Krippendorf for excellent research assistance. The views expressed in this paper are those of the authors alone and do not necessarily represent the views of KfW.

1 Introduction

Today's world is shaped by global challenges including climate change, food and water shortages, rising inequality, and an increasing number of conflicts. The consequences of crises are often particularly felt in Low and Middle Income Countries (LMICs) that have a limited financial capacity to soften their impact. Development finance can be an important remedy and a large and increasing volume of official development assistance (ODA) has been committed over the last decade, reaching a total of USD 211 bn. in 2022 (OECD, 2023).

At the same time, knowledge on the overall effectiveness of such commitments has remained inconclusive, as the evidence available is scarce (Qian, 2015). Certainly, there has been a fundamental and critical advance in the economic analysis of development interventions due to the work of Nobel laureates Banerjee, Duflo, and Kremer, and a corresponding surge in (experimental) impact evaluations (e.g. Olken, 2020). These project-level evaluations are key to understanding individual project effectiveness; and the body of evidence produced to date is impressive. For a complementing assessment of the effects of development assistance efforts, however, there is also interest in looking at development results systematically across projecs and project types, across sectors and countries—in order to understand which implementation factors, program features, and contextual aspects determine (or not) success in delivering intended impacts. This is of relevance given the broad portfolios of typically hundreds of development projects that are simultaneously implemented in any particular LMIC. Only more recently, studies have tried to analyze the role of country- as well as project-characteristics on development outcomes in more detail, mostly using data from the World Bank and the Asian Development Bank (ADB), e.g. Denizer et al. (2013), who initiated this line of systematic research, and Feeny and Vuong (2017), Ashton et al. (2023).

In this paper we analyze the success determinants of KfW Development Banks' projects. KfW manages Germany's development finance commitments on behalf of the German Federal Government, the second largest ODA donor worldwide after the US (OECD, 2023). With a portfolio spanning across all economic sectors and most LMICs, KfW's engagement scope is comparable to that of large multilateral donors. Specifically, we compile a unique database of KfW project evaluations that comprises 5,608 individual success ratings and that is representative of the institution's entire portfolio across 96 partner countries. The analytical sample comprises three types of data sources: (i) key variables coded from the hardcopy evaluation reports, including systematic success ratings using the OECD-DAC criteria in the five dimensions relevance, effectiveness, efficiency, impact, and sustainability; (ii) KfW operational project data; (iii) external

data on economic indicators. Conceptually, the structure of the analytical sample thus corresponds to data constructed for quantitative meta analysis (e.g. Card et al., 2018). Hence, our empirical analysis uses meta regression to correlate success ratings with a host of explanatory project information covering details on project structure, project financing, project complexity, project risks, and context.

Our empirical results provide us with five main substantive conclusions. First, we find a statistically significant positive association between project success rating and financial volume—especially for budget funds, and for higher shares of partner country contributions. This indicates that larger development finance projects with greater ownership are more successful. Second, several dimensions of project complexity display a statistically significant negative correlation with success ratings: longer project durations per se (not delays), longer timespans between mandate—i.e. commitment of funds—and signing of the contract, and technically complex project designs are associated with a less successful rating. Whereas project duration and degree of technical complexity are often difficult to adjust (in particular e.g. in infrastructure projects), a prolonged timespan between mandate and contract could thus serve as an early indicator of a higher risk of project failure.

Third, project risk assessment is a key factor associated with project success. Prior to the start of each project, the potential project risks are forecast by categorizing their severity, occurrence probability and mitigation chances. Our results show that the higher the rate of eventual occurrence of these ex-ante forecast risks, the lower the success rating. This implies that risks are correctly identified, but are difficult to mitigate and/or have not been mitigated sufficiently during implementation. Fourth, almost none of the covariates describing the project structure show a statistically significant correlation with success ratings. This implies, inter alia, that project implementation works equally well with different partner types, and with a varying number of institutions involved. And fifth, our results for the contextual factors indicate a weak positive correlation between GDP growth rates and project success, but no statistically significant role of democracy or fragility indices on success ratings.

In the debate on development project effectiveness, these findings are particularly relevant because they imply that many of the characteristics that matter most for projects delivering on their intended impact are under the influence of donor agencies and partners: project (co-) financing, risk anticipation and risk management, complexity of the design, and partner ownership and integration in the project. Given that research has found project characteristics to correlate similarly with project success across different donors (Bulman et al., 2017; Briggs, 2020), our results are informative for the diverse panorama of donor institutions as well as recipient countries.

Our paper contributes to the literature in four main ways. First, we construct a novel

database with several key features: by coding more than 30—partially new—covariates describing project characteristics across the project life-cycle we provide the most detailed project data to date, to the best of our knowledge. These covariates—e.g. partner type, various measures of complexity and risk—directly address key gaps identified in the literature regarding quality and depth of micro-level information.[1] Second, the granularity of the dependent variable using five dimensions of success ratings—and measured in a systematic way across countries, sectors, and over time—allows us to identify relevant variation in project success. This provides a more nuanced perspective on the achievements of development finance projects, since previous research has had to focus on the overall project rating only. In addition, we can investigate heterogeneous patterns by sector and region.

Third, our data allow us to control for, and probe, evaluator- and evaluation-specific effects. Related research partially relies on self-assigned ratings from project leaders heading the project (Denizer et al., 2013; Feeny and Vuong, 2017; Rommel and Schaudt, 2020; Ashton et al., 2023). In contrast, all success ratings in our data are assigned by an independent evaluation function, and we can empirically test the independence of evaluator/evaluation characteristics and the assigned success ratings.

Fourth, our study takes an in-depth look at bilateral donor contributions, complementing the systematic analyses of project success determinants for multilateral agencies (e.g. Mubila et al., 2000; Denizer et al., 2013; Feeny and Vuong, 2017).[2]. Bilateral and multilateral development aid work differently and this may, in turn, also affect project success. Several studies have demonstrated that bilateral aid, more often than multilateral aid, is used to exert influence over the recipient country (e.g. Dreher et al., 2022a; Fuchs et al., 2014). This could affect the selection of implemented projects and its success determinants differently for bilateral than for multilateral projects for two reasons: First, bilateral donors have a stronger influence on the project and better control over it, which might lead to comparatively better success ratings. Second, with the political dimension playing a role in bilateral development aid, the actual success of the project may not be as important as it is for a multilateral donor, which might lead to comparatively worse success ratings.

In addition to these core contributions, the paper—more generally speaking—is also related to the literature discussing the relationship between individual project or coun-

[1]For example, our variables respond to Bulman et al. (2017), who argue that "[t]his points to the importance of further work to understand the sources of this variation, for example, by systematically measuring the contribution to project success of project implementing agencies within recipient governments.", or to Ashton et al. (2023) who, based on a recent literature review, conclude that "(...) the quality and suitability of project design, have rarely been investigated (...)" and that existing literature has been "(...) concerned mainly with easily observable characteristics like size, duration, and sector".

[2]Wood et al. (2020) analyze Australian bilateral aid, a comparatively small donor. Honig et al. (2022) compile data including several bilateral donors and also KfW, but the scope of project characteristics is considerably smaller given that their analysis focuses on the breadth of information across donors.

try characteristics and project success in detail (e.g. Chauvet et al., 2010; Kilby, 2015). For example, we provide additional empirical results informing the literature on the role of implementing agencies (Shin et al., 2017; Winters, 2019; Marchesi and Masi, 2021). By adding evidence on a previously understudied donor, we also add to the discussion on the comparability of success correlates across donors (Bulman et al., 2017; Briggs, 2020).

The next section provides background information on bilateral development finance and KfW Development Bank. Section 3 describes the data in detail and presents results from descriptive and graphical analyses. Section 4 delineates the estimation strategy, and sections 5 and 6 present and discuss the meta regression results and robustness, respectively. Section 7 concludes.

2 Development finance and project evaluations at KfW Development Bank

2.1 Bilateral development finance

KfW Development Bank[3] handles the majority of Germany's official Financial Cooperation (FC). In 2022, for instance, KfW committed EUR 10.9 bn. (USD 11.5 bn.) to developing countries around the world (KfW, 2022), making it one of the largest bilateral donors worldwide. The funds mainly stem from the German Federal Ministry for Economic Cooperation and Development (BMZ) and finance projects in Africa, Asia, Latin America and the Caribbean, and South-Eastern Europe. Sector-wise, these engagements address all areas from agriculture to water supply. The institution's breadth of engagement is comparable to that of large multilateral development financiers, but its bilateral nature makes it a particular interesting case to study given that bi- and multilateral aid operate differently (Biscaye et al., 2017; Dreher et al., 2022b; Findley et al., 2017; Rommel and Schaudt, 2020).

Funds committed by KfW are implemented via projects that comprise dedicated investments. They are designed jointly with and implemented by local—mostly public—partner agencies such as line ministries, with whom financing agreements are concluded, at times together with international co-financing institutions. This process entails defining the Theory of Change (ToC), outlining the results framework of the development finance project, including target indicators for project performance. Once projects are completed, a completion report is conducted—summarizing the project's

[3]KfW is a German state-owned investment and development bank. Its name originally comes from "Kreditanstalt für Wiederaufbau"—meaning "credit institute for reconstruction".

results from the perspective of KfW project managers. For a representative subset of projects, this is followed by an independent ex-post evaluation of project success (or failure), taking place approximately three years after project completion.

2.2 Project evaluations

At KfW, the independent evaluation function FCE (Financial Cooperation Evaluation) is responsible for carrying out project evaluations. A random sample of 50% of the projects, stratified by nine sectors, is drawn from all completed projects for each year. The sample is, hence, representative for KfW's entire FC portfolio. All ex-post evaluations conducted at KfW adhere to the internationally established OECD-DAC criteria. That is, each evaluation systematically assesses the five criteria (a) relevance, (b) effectiveness, (c) efficiency, (d) impact and (e) sustainability of the given project.[4] Each criterion is rated on a discrete scale from 6 (best) to 1 (worst), i.e. ranging from "very good" to "highly unsatisfactory". Each evaluation also assigns an overall success rating with the same range.[5]

From the annual sample of projects, one third is evaluated by FCE staff, one third by external consultants, and one third by seconded colleagues from KfW's operational departments. The governance of every single evaluation, however, lies with FCE. That is, (i) FCE supervises external consultants and seconded colleagues, (ii) all reports are peer-reviewed internally within FCE, and (iii) the absence of conflicts-of-interest is ensured: Specifically, any person involved in the evaluation process must not have worked on the project or within the responsible department during its implementation. Each evaluation follows a structured process entailing conceptual design, desk study, on-site visit and/or support from a local consultant, and report writing. Summary evaluation reports—with standardized table of contents—are published on KfW's website.

Each evaluation thus constitutes an expert assessment of project success or failure, following an internationally established methodology. Another benefit from exclusively relying on DAC-criteria is that all project evaluations are guided by the same normative framework—independent of regional or thematic focus—addressing concerns that development objectives cannot be compared across sectors (Denizer et al., 2013; Feeny and Vuong, 2017). For the purpose of our study, therefore, the use of DAC-criteria pro-

[4] These five criteria were defined by OECD-DAC in the 1990s. A sixth criterion, "coherence", was only added in 2020, and therefore most evaluations in our sample cover five criteria.

[5] In general, the overall rating is calculated as the rounded, unweighted average of the five criteria ratings. There is one specific exception, however: If one or more of the three criteria sustainability, effectiveness or impact are rated as 3 or below, then the overall project cannot be rated higher than 3, independent of the ratings assigned to the other criteria. This particular scenario applying the so-called *Knock Out-Criteria* concerns 37 of the 1,124 project evaluations, or 3.3% of our sample.

vides us with a large sample of individual success ratings that were systematically and consistently assigned over the entire sampling period.

In contrast to the ADB and World Bank, KfW's evaluation portfolio does not include self-assessments from operational staff, and it is selected on a strictly random basis. Thus, our data are not prone to selection biases (Kilby and Michaelowa, 2019), or to overly favorable ratings assigned by project managers themselves (Bulman et al., 2017; Ashton et al., 2023). Still, even when conducted by a formally independent evaluation function, the autonomy of such bodies can be called into question given that they are based within the institution (Denizer et al., 2013). While the same critique could, in principle, be translated to KfW, several reasons speak against it: First, the director of evaluation is recruited externally from academia, and reports directly and only to the executive board; second, evaluation results are publicly shared; third, the broad set of evaluators (FCE staff, external consultants, seconded operational staff) guarantees the absence of conflicts-of-interest; fourth, FCE's methodology is reviewed by an external body, the German institute for Development Evaluation (DEval). Moreover, we can empirically test the independence of assigned ratings and evaluator characteristics and discuss this in a dedicated part of section 4.

3 Data and descriptive analysis

3.1 Meta sample construction and summary statistics

The sample is constructed from all $N = 1,124$ evaluations of development finance projects that FCE conducted between 2007 and 2021, yielding a sample of $N = 5,608$ observations on project success ratings (five ratings per evaluation). This is, to our knowledge, the most extensive and up to date database on bilateral financial cooperation evaluations worldwide from a single donor.

It covers projects implemented in a total of 96 low and middle income partner countries and is representative for KfW's FC portfolio (cf. Appendix Table A1). As evaluations are conducted after project completion, different project durations imply that our sample effectively contains development projects that started as early as 1990 and as late as 2019. Each of the 1,124 development finance projects in the sample has a unique ID, allowing us to merge system variables from KfW databases with information coded from evaluation reports, covering rich information on project characteristics (the "micro" variables). In addition, we combine these data with external statistics on contextual factors in the countries during the time of project implementation (the "macro" variables). Conceptually, the resulting analytical sample combines three types of data sources—(i) key variables coded from the hardcopy evaluation reports, (ii) KfW op-

erational project databases, (iii) external data on economic indicators—and therefore corresponds to data constructed for quantitative meta analysis (e.g. Card et al., 2018).

3.1.1 Dependent variable: Standardized project success ratings

The main variable of interest is the individual rating assigned for each DAC-criterion on a 6-1 scale for a given project. This allows us to utilize the granularity of the full set of individually assigned ratings instead of only relying on the overall project rating. Appendix Figure A1 displays the respective distribution of each DAC-rating in the data at the project evaluation level ($N = 1,124$ each). The majority of overall success ratings (top left) are either "4" or "5", with more than 400 cases each, while only few evaluations rate projects in the most successful category "6" (44 projects in total). The overall share of ratings with "2" and "3" amounts to 19%, or 204 projects. The overall mean rating is 4.21.

The remaining panels for the five individual criteria indicate several patterns: First, the distributions for "effectiveness" and "impact" are rather similar to the overall rating, each with an average rating of 4.34. Second, "relevance" displays the highest share of successful ratings with "5" and "6", and thus the highest overall mean rating (4.85). Third, both "efficiency" and "sustainability" show slightly less successful average ratings, attaining 4.07 and 4.18, respectively.

The core of our analysis uses the pooled sample of these individual ratings. For robustness and comparability with the related literature, we also construct two additional outcome variables: i) An alternative overall project rating based on an unrounded, unweighted average of all five individual DAC-Criteria ("arithmetic rating"); and ii) a binary variable indicating whether a given project can be considered successful overall. This is indicated by an overall rating of 4, 5 or 6.

3.1.2 Explanatory variables (i): Micro-level project characteristics

At the micro-level, we construct more than 30 variables capturing all dimensions of key project characteristics. Specifically, we can distinguish the four clusters (i) *financing* of the project, (ii) *structure* of the project, (iii) *complexity* of the project, and (iv) *risks* for implementation. Appendix Table A2 presents summary statistics for the main variables within each dimension, along with several macro variables and the distribution by sector.[6] The table shows the full sample (column 1) and a stratification by major regions, i.e. Sub-Saharan Africa (SSA, column 2), Asia/Oceania (3), Europe/Caucasus (4), Latin America and the Caribbean (5), and Middle East and Northern Africa (MENA, column

[6]For a detailed codebook of all variables used, see Appendix Table A8 and A9.

6). Recall that this sample of project evaluations is representative for KfW's development finance portfolio, indicating that the majority of projects are in SSA ($N = 428$ evaluations), followed by Asia/Oceania ($N = 281$).

The "average" development finance project (column 1) has a total volume of EUR 41.7 million, 16% of which are contributed by the country counterpart (top panel on "Financing"). The panel on "Structure" shows that co-financing occurs in 21% of projects on average, varying across regions from 11% (MENA, column 6) to 30% in SSA (column 2). The average number of institutions involved in a development finance project is four. The variable "project manager turnover" relates the total number of project managers in a given project to the project duration in years—implying that at an average of 0.48, the project manager changes every second year.

Looking at project "complexity" (third panel), the average project duration amounts to seven years, ranging by region from 5.7 (Europe/Caucasus) to almost nine years (MENA). These averages relate to the fact that in MENA the execution of projects is delayed in 48% of cases, while this is the case for only 12% in Europe/Caucasus. The overall average share of delayed execution is 23% (column 1). The share of technically complex projects ranges widely from 15% in Latin America to 67% in Asia/Oceania (average: 48%).

Project appraisals identify and specify potential risks for project success ex ante. As the fourth panel (Risk) indicates, the average number of ex ante identified risks is four, with very little variation across regional sample splits. Our data also capture to what extent these risks actually occurred during implementation: 55% of ex-ante identified risks occurred in practice, an average that is somewhat higher in SSA (62%), and somewhat lower in Europe / Caucasus (49%).

3.1.3 Explanatory variables (ii): Macro-level contextual factors

Data on the country context that projects were implemented in are taken from official, publicly accessible databases and merged to our micro-variables using country ISO-codes and information on the project life-cycle: indicators are always measured for the specific country at the specific time the project was implemented. We incorporate four variables in our analyses: GDP p.c. growth, measures of democracy as well as fragility, and total population.[7]

The bottom panel of Table A2 displays the distribution of projects by sector. Some patterns by region are notable: In SSA (column 2), water supply (17%) and health (19%) are major sectors, the latter also being the case in Asia/Oceania (column 3). In

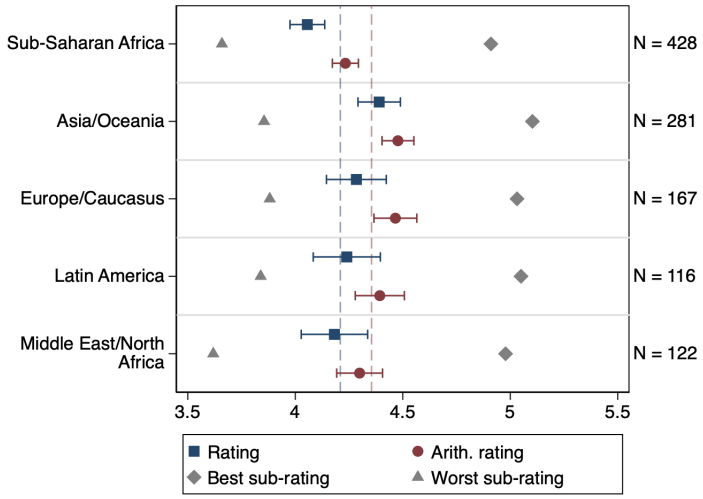
[7]Sections A.3.1 and A.3.2 in the Appendix detail how averages over time and missing observations are computed for these variables.

Europe/Caucasus, water supply (28%) and finance (24%) are the main sectors, while in Latin America agriculture and environment are predominant (34%). In MENA, again water supply plays a major role (32%, column 6).

3.2 Descriptive analysis

Both the country and the sector where projects are allocated constitute two of the most distinctive project characteristics. Indeed, typically donor institutions are institutionally organized along these dimension. This is also the case for KfW Development Bank and reflects how vital this distinction is for project implementation processes. The descriptive analysis therefore continues with a visual inspection of project success patterns by region and sector, respectively, using forest plots. This representation also reflects the meta-analysis nature of our data.

Figure 1: Forest plot of ratings by region



Note: The figure displays mean values of evaluation ratings by region: Blue squares denote average overall ratings (i.e. calculated from the rounded unweighted overall rating assigned to a project in the evaluation), red dots denote average arithmetic ratings (i.e. the unrounded arithmetic mean of the five DAC criteria ratings), diamonds denote means of the highest DAC-ratings per project and triangles denote means of the lowest DAC-ratings per project. 95% confidence intervals illustrated by whiskers. The blue and red dashed lines mark the sample mean of overall and arithmetic, respectively. Observations are weighted by the inverse number of projects evaluated in the corresponding evaluation report. The y-axis on the right hand side gives the number of observations per category. 10 projects implemented in multiple regions excluded.

3.2.1 Overall success rating by region

Figure 1 shows a forest plot of overall success ratings by region. The blue square represents the average *overall rating*, i.e. the rounded (to a full rating) average of the five criteria ratings that is reported in the evaluation report. The red dot represents the average *arithmetic rating*, i.e. the unrounded, unweighted average of the five criteria

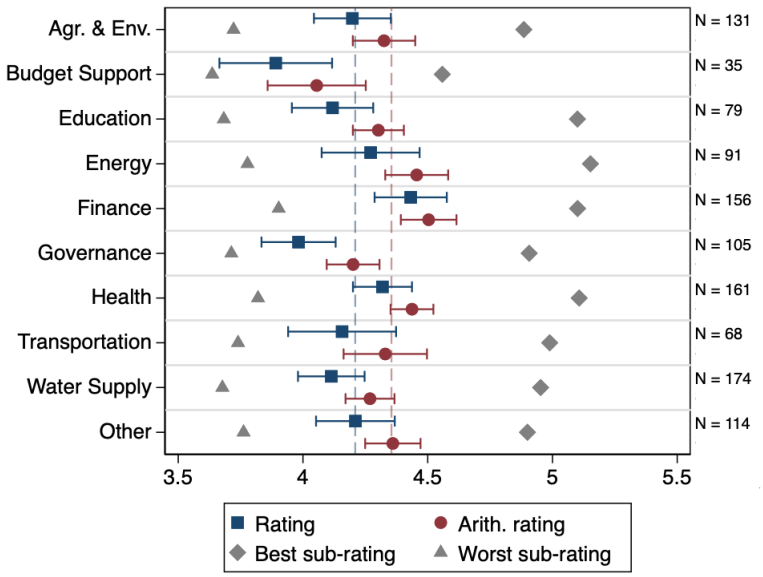
ratings. The respective average of the lowest (triangle) and highest (diamond) DAC-criterion ratings are also shown.

The figure indicates several patterns. First, the overall, rounded rating assigned to the project in the evaluation is always lower than the arithmetic mean of the individual criteria ratings. The difference, however, is not very large (4.21 overall vs. 4.36 arithmetic). Second, the respective regional averages are relatively close to the overall averages rather than widely dispersed. However, third, there are some visible regional differences. In SSA, both means are statistically significant below the overall means (4.06 overall, and 4.23 arithmetic). In Asia/Oceania, on the other hand, the mean ratings are statistically significant above the overall means (4.39 and 4.48, respectively). Finally, in SSA and the Middle-East/North Africa (MENA) the worst sub-rating (triangle) is consistently lower than in the other regions.

3.2.2 Overall success rating by sector

Figure 2 displays the corresponding forest plot of average success ratings by main economic sector. The y-axis on the right indicates the sectoral distribution of project evaluations and reflects the summary statistic shown in Table A2: Inter alia, the three largest sectors are finance, health, and water supply, with a share of 15% each ($N = 156$, $N = 161$, and $N = 174$ evaluations, respectively).

Figure 2: Forest plot of ratings by sector



Note: The figure displays mean values of evaluation ratings by sector. See notes for Figure 1.

The figure illustrates notable variation in project success ratings across sectors: Looking at overall average ratings (i.e. blue squares), finance (4.44) and health (4.32) display

the most successful ratings, the former statistically significant above the overall average. Projects in the energy sector are also comparatively successful and slightly above average (4.27), with relatively wide confidence bands. On the other hand, budget support (3.89) and governance sector (3.98) lie statistically significant below the overall average.

4 Methodology

4.1 Empirical specification

As delineated in the previous section, our meta sample combines rich information on several dimensions of project characteristics with contextual information. Given this structure of our data, we fit the following regression to explain variation in project success:

$$\begin{aligned} Rating_{icrtp} = \alpha\, Fin_{ir} + \beta\, Struct_{ir} + \gamma\, Complex_{ir} + \eta\, Risk_{ir} + \lambda\, Eval_{r} \\ +\theta\, Macro_{cp} + \delta Z'_{irt} + \epsilon_{ct}, \end{aligned} \tag{1}$$

where $Rating_{icrtp}$ denotes the respective DAC-rating dimension of project i, located in country c and evaluated as part of evaluation-report r, written in year t. Fin_{ir}, $Struct_{ir}$, $Complex_{ir}$, $Risk_{ir}$ and $Eval_{r}$ are vectors of relevant project-specific variables capturing the clusters financing, structure, complexity, risks, and evaluation, respectively, while vector $Macro_{cp}$ captures country-specific characteristics at the time of project implementation p. Specific variables within each dimension are discussed further in the results section.

Lastly, $\boldsymbol{Z}_{ir}$ controls for a comprehensive set of additional project-specific variables comprising fixed effects for sector, region, period of implementation as well as evaluation (5-year intervals each, starting in 1990). Robust standard errors are clustered at the country-(c) evaluation-year (t) level. We estimate equation (1) using weighted least squares (WLS), where the weights are given by the inverse number of projects evaluated in the corresponding evaluation report. This approach appropriately reflects the research question and data structure (Denizer et al., 2013; Card et al., 2018).

The main analysis focuses on the pooled sample, using the full set of projects' individual DAC-criteria ratings as outcome variable. The analysis is organized along the key thematic dimensions of interest: That is, we first investigate evaluation features and the independence of assigned ratings (in section 4.2), and then in the results section (section 5) we stepwise introduce and present results for the four project characteris-

tics clusters, as well as for the country context.

Adding to the full sample results, we stratify the sample by region and sector, respectively, to investigate and highlight heterogeneities in project success along these dimensions. From a methodological perspective, several additional analyses and robustness checks are added subsequently: First, we fit equation (1) for each DAC criterion separately to see whether micro and macro variables correlate across these dimensions differently. Second, sensitivity of the outcome variable is verified using ordered probit and probit models (the latter for a success/failure binary indicator). Finally, as a robustness check for the selection of the variables and to reduce the potential of overfitting, we also estimate the model using an adaptive Least Absolute Shrinkage and Selection Operator (LASSO) technique (Zou, 2006). Such an approach reduces the model to the key variables in a first step whilst penalizing large coefficients before the normal WLS model is estimated on the reduced set of variables.

It has to be noted that the coefficients obtained from estimating the model in equation (1) are prone to endogeneity similar to other related research (e.g. Denizer et al., 2013; Ashton et al., 2023). Development aid responds to macro-economic deterioration and political incentives, which are likely to simultaneously affect the observed outcomes. Despite a rather comprehensive and detailed set of project characteristics, we cannot measure all project design features, which in the given context may also respond to unobservable conditions on the ground. Furthermore, finding valid instrumental variables in such settings has proven to be challenging, impeding the identification of causal effects (Bulman et al., 2017; Feeny and Vuong, 2017). When discussing our findings, we therefore point to immediate as well as alternative interpretations that potentially underlie observed estimates. Given that the empirical analysis takes into account an extensive set of fixed effects and control variables, however, we are confident that our results account for unobserved factors to the extent possible, thereby providing interpretable, relevant, and informative results on the determinants of success and failure of development finance projects, in particular in combination with past studies on the topic (Denizer et al., 2013; Feeny and Vuong, 2017; Ashton et al., 2023).

4.2 Independence of ratings

Our analysis benefits from the fact that all success ratings are based on a coherent evaluation methodology. In fact, it is precisely due to the systematic rating framework provided by the DAC-criteria—and applied to 1,124 evaluations over 1.5 decades—that it is possible to construct these data. This coherent, systematic foundation of the data generation process notwithstanding, there is a possibility that other evaluation-specific characteristics may be significantly related to the assigned outcomes due to

potential biases arising in the evaluation process. We test for these concerns in turn, and report the corresponding results in Table 1.

First, the sample time lag between the project completion report (i.e. the formal end of the project) and the evaluation report is 3.27 years. This duration might be structurally related to the success rating, since information for projects assessed later might not be as readily available. Also, certain evaluations may only be conducted with delay due to ongoing conflicts in a given country. Such instances might simultaneously affect the outcome. The first row of Table 1 reports some evidence for such a relationship: In column 1, the coefficient is negative but not significant; when including the entire set of controls, however, the estimate turns significant, indicating that projects assessed at a later stage receive lower ratings, on average (column 5). We therefore control for the time lag between project completion and evaluation in all models.

A second potential bias concerns the type of evaluator. Whereas all evaluations, ultimately, are conducted under the governance and quality assurance mechanisms of FCE, in practice there are four evaluator categories (cf. section 2): FCE staff, seconded colleagues from KfW operational units ("internal"), external evaluation consultants, and internal plus external combined. Ex-ante, it is a theoretical possibility that certain types of evaluators systematically assign, on average, too positive or too negative ratings (e.g. it can be plausibly argued that internal evaluators might be tempted to rate too successfully given their expectation that at some time in the future their own projects will also be evaluated). It is one strength of our data that they contain evaluator type information, allowing us to empirically investigate this potential bias. Rows 2–4 in Table 1 report the results. Both the reduced (columns 2 and 4) and full specifications (column 5) indicate that the magnitude of the point estimates is small and there is no statistically significant correlation between evaluator type and assigned rating.[8] This is a reassuring finding: The unbiasedness of success ratings is not only plausible given the structural independence of the evaluation function and its evaluators, but is in fact an empirical reality.

A third internal process that could potentially influence ratings is the timing of evaluations during the calendar year. Due to the annual sampling process, the evaluation function has the objective to achieve a certain number of evaluations each fiscal year, and the annual count to achieve that number stops on December 31st. This leads to a clustering of evaluation reports at the end of the fiscal year: 30% of the reports in our sample were finalized in December, and 15% in November. The remaining 55% are relatively equally distributed across the other ten months. Whereas the pure number of reports per month is no cause for concern, one might conjecture that last-minute

[8] Apart from this regression framework, a one-way ANOVA test also reveals no statistically significant difference between the four groups and the project rating (not shown, available on request).

reports might potentially be associated with either more positive (in order to finish the report on time) or more negative (the reason why the evaluation took so long) success ratings. Rows 5–15 in Table 1 report the corresponding estimation results and indicate that there is no such pattern recognizable in the data, in particular not concerning any end-of-the-fiscal-year pattern. Only April and January are (marginally) significantly different from the other months, but these are the two months with the lowest number of evaluation reports and simultaneously slightly more positive mean overall ratings (4.45 and 4.47, respectively, the remaining ten months are all in-between 4.08 and 4.44), such that we interpret this as a deviation at random. Nonetheless, we include evaluation month as a control variable in all subsequent specifications.

Table 1: Test for independence of success ratings: correlation with evaluation-specific characteristics

Dep. variable: Rating (Pooled)	(1)	(2)	(3)	(4)	(5)
Time between final review and EPE	−0.013			−0.016	−0.034**
	(0.012)			(0.012)	(0.015)
Evaluation type (Base: FC E):					
-External		−0.046		−0.088	−0.088
		(0.222)		(0.221)	(0.200)
-Internal + external		−0.009		−0.008	−0.019
		(0.073)		(0.074)	(0.070)
-Internal		0.085		0.078	0.078
		(0.075)		(0.074)	(0.070)
Evaluation month (Base: December):					
-January			0.184	0.201	0.206*
			(0.138)	(0.140)	(0.112)
-February			0.044	0.049	0.113
			(0.123)	(0.122)	(0.108)
-March			0.080	0.095	0.080
			(0.098)	(0.099)	(0.103)
-April			0.202**	0.211**	0.271***
			(0.094)	(0.098)	(0.099)
-May			0.014	0.023	−0.008
			(0.110)	(0.109)	(0.103)
-June			−0.019	−0.003	0.046
			(0.108)	(0.109)	(0.114)
-July			−0.012	0.008	0.039
			(0.112)	(0.113)	(0.105)
-August			0.070	0.087	0.003
			(0.090)	(0.089)	(0.092)
-September			−0.099	−0.080	−0.085
			(0.118)	(0.116)	(0.100)
-October			−0.024	−0.023	−0.001
			(0.091)	(0.090)	(0.092)
-November			0.016	0.021	0.025
			(0.082)	(0.081)	(0.078)
Full specification					Yes
Sector and region indicators	Yes	Yes	Yes	Yes	Yes
Sub-rating indicators	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes
Observations	5,608	5,608	5,608	5,608	5,458
Adjusted R^2	0.14	0.14	0.14	0.14	0.23

Note: Table entries are coefficients from WLS regressions with the pooled rating as dependent variable. Weights are given by the inverse of the number of projects evaluated in the corresponding evaluation report. Other controls include the year of project start as well as evaluation year (both 5-year intervals). In the full specification (column 5), all other variables from Tables 3-6 are included. Standard errors in parentheses clustered at the country-evaluation-year-level. Significance at or below 1% (***), 5% (**) and 10 percent (*).

A final issue in which institutional evaluation processes might be correlated with suc-

cess ratings is trends: Over the years general trends toward better projects—and/or even more ambiguously—better ratings could potentially bias our results. In fact, we observe a slight trend towards better ratings over the sample period; however, mean ratings using five-year evaluation completion brackets from 1990 onward are not significantly different from one another (not shown in the table for brevity). Nonetheless, all regressions control for year of evaluation by means of these five-year period indicators.

5 Empirical results

Previous studies have highlighted the importance of project-level factors to explain the success or failure of development projects for multilateral organizations (Denizer et al., 2013; Bulman et al., 2017; Feeny and Vuong, 2017). Our results support and strengthen this result for bilateral development aid. As an initial analytical step in relation to this literature, we calculate the between-country variation in project success and regress country fixed effects on a binary success outcome variable for each year the projects in our sample were active.[9] From the resulting R^2, we derive the share of variation that can be explained by country factors, i.e. the environment in which the projects are implemented. Our result is comparable to that for World Bank projects (Denizer et al., 2013) and indicates that 34% (20% for the pooled sample) of project variation stems from between-country variation. In the following empirical analysis, we thus examine an extensive set of project-level micro variables to contribute towards better understanding determinants of success and failure.

5.1 Cluster (1): Project financing

Whereas previous research had to revert primarily to financial volume as the only proxy for complexity, we are able to address project complexity and project financing separately. Project financing is the first cluster of project-level variables we analyze, providing a nuanced perspective by including information on seven financial variables such as aid type or share of counterpart contributions. The regression results are presented in Table 2. This Table and the subsequent Tables 3–6 are structured as follows: in the first columns, covariates for the relevant cluster are included in the meta regression one by one; and the last column in each table shows coefficients from the full specification, which includes all variables from Tables 1–6.

There is some indication (row 1) that financially larger projects are systematically cor-

[9] See section A.3.3 in the Appendix for a detailed description of the methodology.

related with more successful ratings.[10] The point estimate is positive and statistically significant in the reduced specification (column 1), but becomes insignificant in the full specification (column 9). Financially larger projects may comprise straightforward infrastructure investments or politically prominent showcases receiving more attention, thus making implementation easier.

Table 2: Determinants of success ratings, cluster (1): Project financing

Dep. variable: Rating (Pooled)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Total volume (log)	0.049**							0.029	0.037
	(0.021)							(0.028)	(0.029)
Aid type (Base: Loan):									
–Grant		0.052						0.093	0.105
		(0.088)						(0.098)	(0.087)
% counterpart contributions			0.241**					0.196*	0.145
			(0.112)					(0.117)	(0.118)
Budget funds (log)				0.070**				0.057	0.095**
				(0.029)				(0.037)	(0.042)
% budget funds of ODA					0.000			−0.000	−0.000
					(0.000)			(0.000)	(0.000)
% project funds of GDP						0.000		−0.000	−0.000
						(0.000)		(0.000)	(0.000)
Disbursement vs. commitment							0.169	0.178	0.137
							(0.165)	(0.163)	(0.156)
Full specification									Yes
Sector and region indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sub-rating indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,608	5,608	5,608	5,608	5,608	5,608	5,608	5,608	5,458
Adjusted R^2	0.15	0.14	0.15	0.15	0.14	0.14	0.14	0.15	0.23

Note: Table entries are coefficients from WLS regressions with the pooled rating as dependent variable. % budget funds of ODA and % projects fund of GDP are re-scaled by 1 million. Weights are given by the inverse of the number of projects evaluated in the corresponding evaluation report. Other controls include the year of project start as well as evaluation year (both 5-year intervals) and evaluation month. In the full specification (column 9), all other variables from Tables 1-6 are included. Standard errors in parentheses clustered at the country-evaluation-year-level. Significance at or below 1% (***), 5% (**) and 10% (*).

From a donor perspective, beyond total investment volume, more leverage potentially lies with the budget funds that are committed—i.e. broadly loan vs. grants—as well as the share of counterpart financing contributed by the partner government. KfW staff might, for example, be able to exert higher pressure on contractual and regulatory procedures like due diligence when funds are committed as a loan, possibly resulting in better outcomes. Looking at the results in Table 2, when compared to grants—which represent 90% of development finance projects in the sample—loans do not perform significantly better (row 2).

At the same time, the correlation between the share of counterpart contributions and project success is (marginally) statistically significant and positive (row 3, columns 3 and 8, though not in the full specification in column 9). Intuitively, greater commitment by local partner governments could be expected to be associated with better ratings, such that this finding underlines the importance of ownership as a key principle of development cooperation. The overall tendency of a positive association between variables related to the financial volume of a project and project success that can be

[10] Recall that the total volume refers to the costs of the entire project, i.e. including commitments by the government itself and/or other donors.

taken from Table 2 is further highlighted by a significant positive correlation between budget funds and project ratings in both the reduced and full specification (row 4).

5.2 Cluster (2): Project structure

Details of the project structure are decided at project appraisal and are at the discretion of KfW. Structural design features are, in theory, highly relevant from a policy perspective and could help improve development project effectiveness. There is some evidence that a tailored project design is a determinant for development outcomes on both the individual project- (e.g. Khwaja, 2009) and aggregate-level (e.g. Wane, 2004). This entails, e.g., deciding whether to implement a project along international partners in a co-financing arrangement, which is the case for 21% of projects in our sample. To increase cooperation is a common pledge among donors, largely due to the supposed positive effects attributed to it: More streamlined efforts toward developmental impacts and increased efficiency with regards to disbursement conditions have been affirmed in both the Paris Declaration and the Accra Agenda (OECD, 2022b). Our results provide only limited support for this hypothesis, as the coefficient on co-financing arrangements in Table 3 is positive and at the margin of significance (row 1, column 1, t-value 1.65).

Development finance projects are often implemented along with technical assistance to support local partner agencies (27% of projects in the sample). A plausible prior belief is that these measures are associated with improved project outcomes. However, the direction is not straight-forward, as it could be particularly weak partners who receive such support in the first place. Such negative selection bias has been argued for example to influence the relationship between more diligent project preparation time and unfavorable ratings (Denizer et al., 2013). The estimation results for accompanying measures in Table 3 are not statistically different from zero (row 2, columns 2, 8, and 9), a result that does not allow to disentangle the role that these measures play or not. In fact, the insignificant point estimate could indicate that, on average, successful accompanying measures mitigate the negative selection effect.

Several more structural design features are worth considering: For instance, donors work with a multitude of local implementing partners, yet existing research cannot provide detailed insights regarding these agencies' capacities. Increasingly, projects are implemented with non-state actors, responding to the recognition that governmental partners' capacity is limited (Feeny and de Silva, 2012), and potentially allowing for more participatory development partnerships with the civil society. In fact, such projects have been shown to perform better in some instances (Shin et al., 2017). While certain sectors such as micro-finance are already dominated by private agencies,

in our sample most implementing partners—around 68%—are governmental institutions. Distinguishing different agency types, rows 3–6 in Table 3 find no significant relationship between any of these types and corresponding project success. This is an informative empirical finding for future project design: Agency type is not a key factor for project success, and neither is whether previous cooperation existed (row 7) nor the number of institutions involved (row 8).

Table 3: Determinants of success ratings, cluster (2): Project structure

Dep. variable: Rating (Pooled)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Co-financing	0.091							0.075	0.002
	(0.055)							(0.056)	(0.064)
Accompanying measure		−0.066						−0.048	−0.015
		(0.058)						(0.058)	(0.056)
Agency type (Base: NGO):									
-Mixed			−0.034					−0.032	−0.099
			(0.126)					(0.128)	(0.130)
-Multilateral			0.113					0.080	−0.009
			(0.127)					(0.127)	(0.131)
-Private sector			0.031					−0.008	0.006
			(0.138)					(0.139)	(0.139)
-Government			−0.052					−0.056	−0.101
			(0.105)					(0.107)	(0.107)
Previous cooperation				0.079				0.074	0.066
				(0.053)				(0.053)	(0.051)
Number of institutions					0.008			0.005	0.005
					(0.009)			(0.010)	(0.009)
Project manager turnover						0.402*		0.348*	0.328
						(0.216)		(0.209)	(0.248)
Country office							−0.018	−0.008	−0.043
							(0.053)	(0.053)	(0.056)
Full specification									Yes
Sector and region indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sub-rating indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,608	5,608	5,608	5,608	5,608	5,608	5,608	5,608	5,458
Adjusted R^2	0.15	0.15	0.15	0.15	0.14	0.15	0.14	0.15	0.23

Note: Table entries are coefficients from WLS regressions with the pooled rating as dependent variable. Observations are weighted by the inverse of the number of projects evaluated in the corresponding evaluation report. Other controls include the year of project start as well as evaluation year (both 5-year intervals) and evaluation month. In the full specification, all other variables from Tables 1-6 are included (column 9). Standard errors in parentheses clustered at the country-evaluation-year-level. Significance at or below 1% (***), 5% (**) and 10% (*).

In row 9 of Table 3 we estimate the role of project manager turnover. Project managers are in charge of the team at KfW and are the focal point for all interactions with the partner country and project implementing unit. Their importance for the success of a project is key, and therefore greater turnover could lead to knowledge loss and thus lower ratings (Ashton et al., 2023).[11] Our results indicate an—at first sight—counter-intuitive, positive relationship between the number of project managers per year and evaluation outcomes (reduced specification, column 6). Once we control for all other factors in the full specification (column 9), however, this association is no longer statistically significant. Finally, as a last hypothesis concerning this cluster of project variables, we explore whether a local KfW office in the project implementing country supports the success of a project. The assumption behind this is that such office presence might translate into higher engagement and knowledge in the partner country, result-

[11] Since projects have different durations, we normalize the number of project managers per operational year. A value of one therefore indicates that a project was managed by a new manager during each year when it was operational.

ing in more successful projects (Honig, 2020). The estimation results do not support this hypothesis.

5.3 Cluster (3): Project complexity

The design and, in particular, the implementation of development finance projects is often complex and challenging. Our meta sample allows us to investigate in more detail five features of this complexity. The overall finding from the corresponding results presented in Table 4 is that more complex projects have a lower likelihood of success.

In particular, the first dimension of project complexity captures the duration of the project. As row 1 of the table shows for all specifications (columns 1, 6, and 7, respectively), a longer project duration is strongly and significantly correlated with worse success ratings. The eventual duration of a project has both an implementational component, e.g. delays in contracting or executing, and a structural component, as it also depends on the sector or region where it is placed, which in turn also influence outcomes as described in 3.2. Row 2 of the table specifically investigates the role of delays, and shows that these are not a significant explanation of lower project ratings.

Table 4: Determinants of success ratings, cluster (3): Project complexity

Dep. variable: Rating (Pooled)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Project duration (log)	−0.234***					−0.215***	−0.149**
	(0.069)					(0.073)	(0.075)
Delay		−0.032				0.009	0.009
		(0.069)				(0.070)	(0.069)
Revised ToC			−0.071			−0.060	−0.048
			(0.049)			(0.049)	(0.047)
Years mandate to contract				−0.030		−0.026	−0.048*
				(0.024)		(0.023)	(0.027)
Technical complexity					−0.117**	−0.083	−0.130**
					(0.055)	(0.056)	(0.055)
Full specification							Yes
Sector and region indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sub-rating indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,608	5,608	5,608	5,598	5,608	5,598	5,458
Adjusted R^2	0.15	0.14	0.15	0.14	0.15	0.16	0.23

Note: Table entries are coefficients from WLS regressions with the pooled rating as dependent variable. Weights are given by the inverse of the number of projects evaluated in the corresponding evaluation report. Other controls include the year of project start as well as evaluation year (both 5-year intervals) and evaluation month. In the full specification (column 7), all other variables from Tables 1-6 are included. Standard errors in parentheses clustered at the country-evaluation-year-level. Significance at or below 1% (***), 5% (**) and 10% (*).

Concerning another, related factor of complexity, the length of time between the official commitment of governmental funds and their translation into actual projects as part of a contract is theoretically ambiguous. While a longer time-length from mandate to contract could be an early indicator for eventually hard-to-manage projects, they could also fare better due to thorough preparation (Deininger et al., 1998; Bulman et al., 2017; Kilby, 2015). Row 4 of Table 4 depicts some evidence for the former hypothesis, as the coefficient for the full specification (column 7) indicates a negative, marginally

significant correlation between the length of time from mandate to contract and the success rating.

Additionally, we consider whether the ToC (Theory of Change) outlined at the time of project appraisal was adjusted as part of the evaluation. A change could indicate that the project framework was not adequate in the first place or had to be updated to reflect operational adjustments, hinting towards increased complexity and thus potentially lower ratings (Blanc et al., 2016). However, we find this measure to be irrelevant for the rating obtained. As the last factor in this cluster, we analyze whether technically complex projects are correlated with better or worse evaluation ratings. This "Technical complexity" is a binary indicator variable taking on the value of one if the project required the support of a specific technical advisor, e.g. engineers for infrastructure projects. Row 5 of Table 4 shows that indeed technically complex projects—even when controlling for sector fixed effects—are significantly correlated with less successful project ratings.

5.4 Cluster (4): Project risks

A particularly interesting cluster of micro variables in our data is KfW's internal risk assessment information. Specifically, the data contain information on (i) the number of risks that were identified ex-ante (i.e. before project start), (ii) the percentage of these that actually materialized during project implementation, (iii) the severity of the *overall* risk to project success ex-ante (low/medium/high), and (iv) the expected level of controllability of that overall risk (low/medium/high).

Row 1 of Table 5 presents estimation results for the number of risks identified ex ante. In theory, a larger number implies a more challenging project, yet could also mean that the design is more deliberately thought through to cope with uncertainties during implementation. The results indicate no correlation between the pure number of identified risks and average project success. The key factor that matters for project success, however, is whether and at what rate these pre-identified risks actually materialize: Row 2 consistently shows a strong and statistically significant negative correlation between the share of risks that occurred and the success rating (columns 3, 5, and 6). In fact, the point estimate for the full specification (column 6) implies that projects for which all risks materialize are rated 0.5 points lower. This is a considerable effect size.

Furthermore, rows 3–5 of the table show that high-risk and medium-risk projects are statistically significantly associated with a lower success rating, relative to low-risk projects. Again, the effect is sizable (full specification, column 6): -0.35 rating points on average for high-risk projects, and -0.2 rating points for medium-risk projects, relative to low-risk project. Whether any of these risks was deemed controllable or not ex ante

Table 5: Determinants of success ratings, cluster (4): Project risks

Dep. variable: Rating (Pooled)	(1)	(2)	(3)	(4)	(5)	(6)
Number ex-ante identified risks	−0.004				0.002	0.001
	(0.013)				(0.013)	(0.013)
% ex-ante identified risks occured		−0.504***			−0.464***	−0.486***
		(0.068)			(0.067)	(0.067)
Overall risk (base: low)						
-Medium			−0.251***		−0.185**	−0.203**
			(0.080)		(0.078)	(0.082)
-(Very) high			−0.460***		−0.326***	−0.352***
			(0.086)		(0.084)	(0.088)
-Not assigned			−0.285***		−0.159	−0.219*
			(0.109)		(0.116)	(0.116)
Overall risk control (base: low)						
-Medium				0.075	0.070	0.084
				(0.058)	(0.055)	(0.058)
-High				0.001	−0.065	−0.061
				(0.253)	(0.193)	(0.169)
-Not assigned				0.090	–	–
				(0.096)	(.)	(.)
Full specification						Yes
Sector and region indicators	Yes	Yes	Yes	Yes	Yes	Yes
Sub-rating indicators	Yes	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,608	5,608	5,608	5,608	5,608	5,458
Adjusted R^2	0.14	0.18	0.16	0.15	0.19	0.23

Note: Table entries are coefficients from WLS regressions with the pooled rating as dependent variable. Weights are given by the inverse of the number of projects evaluated in the corresponding evaluation report. Other controls include the year of project start as well as evaluation year (both 5-year intervals) and evaluation month. In the full specification (column 6), all other variables from Tables 1-6 are included. The risk control category "not assigned" is omitted due to collinearity in column 5–6. Standard errors in parentheses clustered at the country-evaluation-year-level. Significance at or below 1% (***), 5% (**) and 10% (*).

does not affect success ratings (rows 6–8 of the table). Overall, these findings imply that relevant risks are correctly identified, but are difficult to mitigate—and/or have not been mitigated sufficiently—during project implementation.

5.5 Contextual variables

Historically, macroeconomic outcomes such as GDP have shaped the discussion around the success of development aid (Isham and Kaufmann, 1999; Qian, 2015). However, development projects ultimately are not only supposed to fuel development, but they are simultaneously affected by the economic environment in which they operate. This holds particularly for GDP growth, the most immediate variable measuring the general economic environment and shocks. The related literature has shown that an environment conducive to growth is a significant predictor for project success (Denizer et al., 2013) —and this is an empirical relationship we also observe in most of our specifications as shown in row 1, columns 1 and 5 of Table 6.

The role of civil liberties and citizen freedom is theoretically more ambiguous: Policies in democracies could be more aligned with citizens' needs than in autocracies, yet the latter might provide a more stable institutional environment. Indeed, the literature has found conflicting relationships for World Bank and ADB financed projects (Isham et al., 1997; Feeny and Vuong, 2017). We correlate Freedom House Democracy scores with success ratings, however cannot confirm previous results in either direction (row

2). In light of donor-targeting decisions partially based on governance criteria (Feeny and Vuong, 2017), this is highly relevant. This particularly holds for German bilateral aid, which has recently put more emphasis on good governance criteria in commitment decisions as part of its reform partnerships (BMZ, 2022).

The institutional environment plays a crucial role for the success of aid interventions, particularly because most projects are implemented jointly with governmental partners. A reasonable expectation is that in conflict-prone countries, i.e. where state fragility is more pronounced and institutional quality lower, it is more difficult for projects to deliver on their objectives (Caselli et al., 2021). For example, World Bank projects have been shown to be more fruitful in post-conflict settings with sustained peace (Chauvet et al., 2010). Using the State Fragility Index—incorporating measures of governance effectiveness and legitimacy—we find no statistically significant relationship, however (row 3).

The size of a country in terms of population is potentially adversely related to the probability of success, given the growing complexity with governing more people (Feeny and Vuong, 2017). We find no evidence for this in our meta sample either (row 4). Lastly, contextual factors beyond the country-level that are not specific to projects and vary over the period of implementation likely also matter. An example would be institutional arrangements among donors that increase delivery on projected outcomes. While we cannot account for those directly, we include indicators for five-year brackets of the year of project appraisal, capturing changes in institutional arrangements over time.

Table 6: Determinants of success ratings: Country context

Dep. variable: Rating (Pooled)	(1)	(2)	(3)	(4)	(5)	(6)
GDP p.c. growth (annual)	0.017**				0.016*	0.011
	(0.008)				(0.009)	(0.008)
Freedom House Democracy score		0.000			−0.006	−0.018
		(0.018)			(0.021)	(0.021)
State Fragility Index			−0.008		−0.008	−0.006
			(0.006)		(0.007)	(0.008)
Population (log)				−0.002	0.002	−0.029
				(0.017)	(0.017)	(0.022)
Full specification						Yes
Sector and region indicators	Yes	Yes	Yes	Yes	Yes	Yes
Sub-rating indicators	Yes	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,608	5,468	5,468	5,608	5,468	5,458
Adjusted R^2	0.15	0.14	0.15	0.14	0.15	0.23

Note: Table entries are coefficients from WLS regressions with the pooled rating as dependent variable. Weights are given by the inverse of the number of projects evaluated in the corresponding evaluation report. Other controls include the year of project start as well as evaluation year (both 5-year intervals) and evaluation month. In the full specification (column 6), all other variables from Tables 1-4 are included. Standard errors in parentheses clustered at the country-evaluation-year-level. Significance at or below 1% (***), 5% (**) and 10% (*).

6 Heterogeneity and robustness

Disaggregating the results potentially yields further insights and can unmask different patterns within the sample. In addition to the findings for the pooled sample, we stratify the analysis by region and sector, and fit separate regressions for the individual DAC criteria.

6.1 Empirical results by region

Development institutions regularly identify striking differences in projects' success depending on the region where the project was implemented, with Sub-Saharan Africa (SSA) often providing the most challenging environment. Appendix Table A3 disaggregates empirical results by region and shows that indeed the correlation between the various determining factors and the project success rating is heterogeneous. In particular, there are only few variables that play the same significant role throughout all regions: One example of these are the variables in the project risk cluster.

Looking at the regions in turn, in the SSA-sample shown in column 2 two variables stand out: Projects financed via grants fare considerably better than loans, potentially explained by the fact that these instruments are used particularly for fundamental public services such as water supply, where ownership could thus be higher. At the same time, projects led by governmental agencies are significantly rated worse as compared to NGO-led ones. As the flip side of the ownership argument, it could hint towards public institutions' limited capacity when it comes to providing basic infrastructure. Turning to Asia in column 3, higher budget funds are significantly associated with better success ratings. In this region, the country-context appears to matter more than elsewhere. While a positive relationship with GDP p.c. is intuitive, it runs counter to comparable findings for state fragility (e.g. Chauvet et al., 2010). In contrast, larger population size is significantly associated with lower success outcomes.

Micro-variables are more often significantly related with outcomes in Europe (column 4). In particular, covariates in the clusters project structure and project complexity correlate negatively with outcomes, implying that KfW would have a higher leverage to address underlying obstacles ex-ante and during implementation. While for example the number of institutions involved appears to make projects over-complex, at the same time the projects' outcomes are less strongly affected by ex-ante identified risks that eventually materialized. Also, co-financing contributes to more successful outcomes in European partner countries. A puzzling result is that greater democracy is negatively related to project success in our data for this region, which adds to the already ambiguous results found in the related literature (e.g. Isham et al., 1997; Kosack,

2003).

Looking at independent variable clusters across regions, several things are notable in Table A3. First, the point estimate for the total volume is positive almost everywhere, and particularly large in MENA (column 6, financing cluster). Second, in the project structure cluster, co-financing is significantly positively correlated with project success in Europe/Caucasus (column 4), but negatively in MENA (cloumn 6). Third, complexity as driven by project duration is a key challenge in Europe/Caucasus and in Latin America (columns 4 and 5, both significantly negative), while technical complexity is a key challenge in SSA and again in Latin America (colums 2 and 5). Fourth, the one coherent pattern across all regions concerns project risks: While the number of pre-identified risks per se is not significant, the share of risks that materialize is a significant determinant of project success, or failure, in all regions.

6.2 Empirical results by sector

In a next step we conduct the sub-sample analysis on the projects' main thematic focus, as shown in Appendix Table A4. While many of the patterns identified for the pooled sample and the regional stratification are visible in the sectoral results, too, several additional results warrant attention.

First, in the energy sector (column 4), projects with more institutions involved are associated with weaker outcomes, possibly due to undue complexity in a sphere dominated by large-scale utility companies on the partner-side. Longer preparation times before contract closings however relate positively with ratings, potentially hinting towards the role of due diligence in these mostly large-scale infrastructure investments. Two variables are noteworthy for governance interventions (column 7): Against the pattern in most sectors, more institutions involved in the implementation appear to yield better outcomes. Due to the complexity of these projects, a holistic approach might therefore be beneficial for this sector. Similarly surprising, these projects fare better in more fragile contexts, where governance might have already been weak in the first place. On the contrary, fragility is negatively related with agriculture-related projects (column 2). In this sector, larger projects with greater counterpart contribution shares—thus potentially inducing more ownership on part of the partners—are also rated better on average.

For transport-themed projects (column 8), the number of ex-ante identified risks stands out: Its negative relationship with evaluation ratings raises the question how well institutions can mitigate these risks that were apparent before project inception. The identification is a key component of any due diligence, yet the ex-post perspective suggests that investing in projects with large uncertainties should potentially be scru-

tinized more thoroughly in the first place. Lastly, water-sector projects (column 9) are the only ones significantly related to an ex-post revised ToC.

6.3 Individual DAC-criteria

In the next analytical step, we estimate our main specification for five DAC-criteria and the overall rating separately and present results in Appendix Table A5. Each criterion addresses a unique dimension of project success and thus provides an additional, detailed perspective on relevant success determinants. While the focus of *relevance* is on the project layout at the time of inception when adjustments are still viable, *efficiency*, *effectiveness* and *impact* evaluate actual outcomes during implementation. Lastly, *sustainability* concerns outcomes observed at the time of the evaluation, taking into account potential future scenarios of project outcomes. Across the criteria, a first glance reveals that heterogeneities found in the preceding stratifications do not necessarily mirror those at the single criterion level. Nevertheless, only one explanatory variable is consistently significant—the share of eventuated risks—and the other correlates vary considerably.

Given that the success criterion *relevance* (column 2 of Appendix Table A5) reflects project design and its ability to address developmental challenges, the *project structure* variables (panel 2) are of particular interest. However, we find that none of the micro variables in this cluster are significantly related to the rating. This includes financing variables (panel 1) which are still partially—as in the case of the budget funds committed—at the scrutiny of KfW. Looking at determinants in the clusters *complexity* and *risks* (panels 3 and 4), the share of risks that materialized displays a significant negative relationship with relevance. Due to the, in theory, structural disconnectedness over time—ex-ante relevance of project design vs. actual operational risks materializing—this relationship is somewhat surprising and suggests a level of risk tolerance: KfW correctly anticipates operational risks at the time of appraisal, but from an evaluative point of view these may have already been (partially) rooted in the project design itself. This interpretation is corroborated by the (marginally) significant negative relationship between the ex-post adjusted indicators of the Theory of Change (panel 3) and the *relevance* outcome.

With regards to *efficiency* (column 3), projects with greater budget funds appear to fare better (panel 1). This potentially stems from large-scale infrastructure investments that undergo more extensive cost-benefit analyses on part of KfW than projects with regionally spread, small-scale investments. Previous cooperation displays a (marginally) positive influence on project efficiency (panel 2), while technical complexity seems to be detrimental (panel 3). Risk variables as determinants of the efficiency rating show

the same pronounced and consistent empirical pattern we observe across all OECD DAC criteria (panel 4), and the magnitude of the point estimates is even slightly larger for *efficiency*: both the share of ex-ante identified risks that occur and the overall project risk level correlate strongly with lower success ratings. Finally, panel 5 illustrates that the efficiency criterion—closely related to the rate of economic return—is positively associated with the country macro environment, particularly GDP (Isham and Kaufmann, 1999).

When assessing the *effectiveness* of interventions (column 4), in addition to the risk pattern found for all outcome variables, the relationship with delays before actual implementation (variable "years mandate to contract" in panel 3)—i.e. time between intergovernmental agreements and the actual project financing contract—is negative. Potentially, this could already constitute a red flag for later implementation challenges e.g. due to partner capacity constraints. However, the relationship does not materialize further down the project logic as displayed in column 5 for the longer-term *impact* criterion. Here, again technical complexity seems to be the hampering factor. At the same time, concerning both *effectiveness* and *impact*, projects with more budget funds are associated with more successful developmental results.

Lastly, the positive significant determinants of project *sustainability* (column 6) are larger total investments (panel 1) and previous cooperation with project implementing agencies (panel 2). This is a key result concerning two main project design parameters. Another main project parameter, the actual duration from signing of the contract until final review, shows a statistically significant negative relationship with the sustainability rating (panel 3). This finding indicates that overly long project implementation periods already hint at difficulties in attaining project objectives, which then in the longer run translate into lower levels of sustainable results. In addition, at the country-level (panel 5), projects in more fragile environments systematically attain less successful sustainability ratings, a result that adds to similar findings in the literature (Chauvet et al., 2010).

6.4 Robustness

In our main specification, the outcome variable represents individual DAC-ratings that are assigned on an ordinal scale. While estimating the models using WLS allows for straight-forward interpretation of the coefficients, we also estimate equation 1 in alternative specifications to assess the robustness of the coefficients: Appendix Table A6 displays results from OLS and ordered probit models for the pooled ratings (columns 1-3), the overall project rating (columns 4 and 5), the arithmetic rating (columns 6 and 7) and a binary success measure (cf. section 3, column 8) as outcome variables. By

and large, we find that the coefficients (and significance levels) are comparable across the specifications, and in line with our main results from the preferred specification. Some of the key patterns identified above appear more pronounced for the simple binary success indicator (overall grades 4-6 = "success", grades 1-3 *failure"), and it also indicates that larger shares of actual disbursement vs. initial commitments determine more successful projects (column 8).

Our selection of the explanatory variables is based on existing theories and past studies. Yet, it may still be the case that we ignore important, additional variables or put too much emphasis on those variables included. In order to address this issue, we apply an adaptive LASSO approach. This method automates the variable selection and strips the model to its most predictive variables. More precisely, the variables with the most explanatory power are identified, before equation 1 is re-estimated with the thereby identified reduced set of variables. Results are presented in Appendix Table A7. After identifying the variables with the most explanatory power, several control and four main variables in our main specification are dropped. Those are the share of counterpart contributions, indicators for co-financing and accompanying measures, as well as the delay variable. Re-estimating the model with the reduced number of variables mainly confirms the previous results, as the coefficient size for most variables remains similar. However, two additional variables turn statistically significant: The total volume and the number of project managers in a given project, implying that a greater project manager turnover increases the rating. The reduction of variables goes along with a reduced adjusted R^2 of 0.19 compared to 0.23 in the WLS regression.

7 Conclusion

This paper presents a systematic, quantitative analysis of the determinants of success—and failure—of three decades of German bilateral development finance. We construct a unique meta database comprising 5,608 project evaluation ratings, and covering 96 LMICs and all economic sectors. We can empirically test, and establish, the independence of success ratings and evaluator/evaluation characteristics. This is the most comprehensive and up-to-date database on bilateral development finance results worldwide from a single donor. As we are able to include in our meta sample extensive and novel data on project characteristics, thereby addressing key gaps identified in the literature, our analysis yields new insights on the question of what works in development finance.

Specifically, we draw five main substantive conclusions. First, we find a statistically significant positive association between project success rating and financial volume—especially for budget funds, and for higher shares of partner country contributions.

This indicates that larger development finance projects with greater ownership are more successful. In particular, larger financial project volumes positively affect the *sustainability* success rating.

Second, several dimensions of project complexity display a statistically significant negative correlation with success ratings: longer project durations (i.e. implementation durations from contract signing until final review), longer timespans between mandate—i.e. commitment of funds—and signing of the contract, and technically complex project designs are associated with a less successful rating. Whereas project duration and degree of technical complexity are often difficult to adjust (in particular e.g. in infrastructure projects), a prolonged timespan between mandate and contract could thus serve as an early indicator of a higher risk of project failure.

Third, project risk assessment is a key factor associated with project success. In fact, the covariates in the *risk* cluster display the most pronounced, consistent pattern across all specifications and success indicators: the higher the rate of eventual occurrence of the ex-ante forecast risks, the lower the success rating. And the higher the overall risk level ex-ante, the lower the success rating. This implies that risks are correctly identified, but are difficult to mitigate and/or have not been mitigated sufficiently during implementation.

Fourth, almost none of the covariates describing the project structure show a statistically significant correlation with success ratings. This implies, inter alia, that project implementation works equally well with different partner types, and with a varying number of institutions involved. One additional detailed result on project structure is that if a project continues a previous collaboration with an implementing partner, then this is significantly associated with better *efficiency* and *sustainability* ratings.

And fifth, our results for the contextual factors indicate a weak positive correlation between GDP growth rates and project success, but no statistically significant role of democracy or fragility indices on overall success ratings. One detailed, and intuitive, significant result, however, indicates that successful *sustainability* ratings are more difficult to achieve in more fragile contexts.

References

Ashton, L., J. Friedman, D. Goldemberg, M. Z. Hussain, T. Kenyon, A. Khan, and M. Zhou (2023). A puzzle with missing pieces: Explaining the effectiveness of World Bank development projects. *The World Bank Research Observer 38*(1), 115–146.

Biscaye, P. E., T. W. Reynolds, and C. L. Anderson (2017). Relative effectiveness of bilateral and multilateral aid on development outcomes. *Review of Development Economics 21*(4), 1425–1447.

Blanc, M., T. Esmail, C. Mascarell, and J. R. Rodriguez (2016). Predicting project outcomes: A simple methodology for predictions based on project ratings. *The World Bank Policy Research Working Paper No. 7800*. `https://ssrn.com/abstract=2836548`.

BMZ (2022). BMZ reform partnerships. `https://www.bmz.de/resource/blob/86828/3357dcbd9969cb774b6fdeb7dfd75861/marshall-plan-review-outlook-4-years-ba-data.pdf`.

Briggs, R. C. (2020). Results from single-donor analyses of project aid success seem to generalize pretty well across donors. *The Review of International Organizations 15*(4), 947–963.

Bulman, D., W. Kolkma, and A. Kraay (2017). Good countries or good projects? comparing macro and micro correlates of world bank and asian development bank project performance. *The Review of International Organizations 12*, 335–363.

Card, D., J. Kluve, and A. Weber (2018). What works? a meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association 16*(3), 894–931.

Caselli, F. G., A. F. Presbitero, R. Chami, R. Espinoza, and P. Montiel (2021). Aid effectiveness in fragile states. In R. Chami, R. Espinoza, and P. J. Montiel (Eds.), *Macroeconomic Policy in Fragile States*, pp. 493–520. Oxford University Press.

Chauvet, L., P. Collier, and M. Duponchel (2010). What explains aid project success in post-conflict situations? *The World Bank Policy Research Working Paper Series No. 5418*. `https://doi.org/10.1596/1813-9450-5418`.

Deininger, K., L. Squire, and S. Basu (1998). Does economic analysis improve the quality of foreign assistance? *The World Bank Economic Review 12*(3), 385–418.

Denizer, C., D. Kaufmann, and A. Kraay (2013). Good countries or good projects? Macro and micro correlates of World Bank project performance. *Journal of Development Economics 105*, 288–302.

Dreher, A., V. Lang, B. P. Rosendorff, and J. R. Vreeland (2022a). Bilateral or multilateral? international financial flows and the dirty-work hypothesis. *Journal of Politics 84*(4), 1932–1946.

Dreher, A., V. Lang, B. P. Rosendorff, and J. R. Vreeland (2022b). Bilateral or multilateral? International financial flows and the dirty-work hypothesis. *The Journal of Politics 84*(4), 1932–1946.

Feeny, S. and A. de Silva (2012). Measuring absorptive capacity constraints to foreign aid. *Economic Modelling 29*(3), 725–733.

Feeny, S. and V. Vuong (2017). Explaining aid project and program success: Findings from Asian Development Bank interventions. *World Development 90*, 329–343.

Findley, M. G., H. V. Milner, and D. L. Nielson (2017). The choice among aid donors: The effects of multilateral vs. bilateral aid on recipient behavioral support. *The Review of International Organizations 12*, 307–334.

Freedom House (2021). Freedom in the World Data. `https://freedomhouse.org/report/freedom-world#Data`.

Fuchs, A., A. Dreher, and P. Nunnenkamp (2014). Determinants of donor generosity: A survey of the aid budget literature. *World Development 56*, 172–199.

Honig, D. (2020). Information, power, and location: World Bank staff decentralization and aid project success. *Governance 33*(4), 749–769.

Honig, D., R. Lall, and B. C. Parks (2022). When does transparency improve institutional performance? Evidence from 20,000 projects in 183 countries. *American Journal of Political Science*.

Integrated Network for Societal Conflict Research (INSCR) (2018). State fragility index and matrix, time-series data, 1995-2018. `https://www.systemicpeace.org/inscrdata.html`.

Isham, J. and D. Kaufmann (1999, 02). The forgotten rationale for policy reform: The productivity of investment projects. *The Quarterly Journal of Economics 114*(1), 149–184.

Isham, J., D. Kaufmann, and L. H. Pritchett (1997). Civil liberties, democracy, and the performance of government projects. *The World Bank Economic Review 11*(2), 219–242.

KfW (2022). KfW annual review 2021. Retrieved October 30, 2022, from `https://www.kfw.de/About-KfW/Newsroom/Latest-News/Pressemitteilungen-Details_703296.html`.

Khwaja, A. I. (2009). Can good projects succeed in bad communities? *Journal of Public Economics 93*(7-8), 899–916.

Kilby, C. (2015). Assessing the impact of World Bank preparation on project outcomes. *Journal of Development Economics 115*, 111–123.

Kilby, C. and K. Michaelowa (2019). What influences World Bank project evaluations? In N. Dutta and C. R. Williamson (Eds.), *Lessons on foreign aid and economic development: Micro and macro perspectives*, pp. 109–150. Springer.

Kosack, S. (2003). Effective aid: How democracy allows development aid to improve the quality of life. *World Development 31*(1), 1–22.

Marchesi, S. and T. Masi (2021). Delegation of implementation in project aid. *The Review of International Organizations 16*, 655–687.

Mubila, M. M., C. Lufumpa, and S. Kayizzi-Mugerwa (2000). A statistical analysis of determinants of project success: Examples from the african development bank. Economic Research Paper No. 56, African Development Bank, Abidjan, Côte d'Ivoire. https://www.afdb.org/fileadmin/uploads/afdb/Documents/Publications/00157646-FR-ERP-56.PDF.

OECD (2022a). Aid (ODA) commitments to countries and regions [DAC3a]. `https://stats.oecd.org/` (accessed on 24 March 2022).

OECD (2022b). The high level fora on aid effectiveness: A history. Retrieved October 24, 2022, from `https://www.oecd.org/dac/effectiveness/thehighlevelforaonaideffectivenessahistory.htm`.

OECD (2023). Net oda. `https://doi.org/https://doi.org/10.1787/33346549-en/` (accessed on 22 November 2022).

Olken, B. A. (2020). Banerjee, duflo, kremer, and the rise of modern development economics. *The Scandinavian Journal of Economics 122*(3), 853–878.

Qian, N. (2015). Making progress on foreign aid. *Annual Review of Economics 7*(1), 277–308.

Rommel, T. and P. Schaudt (2020). First impressions: How leader changes affect bilateral aid. *Journal of Public Economics 185*, 104107.

Shin, W., Y. Kim, and H.-S. Sohn (2017). Do different implementing partnerships lead to different project outcomes? Evidence from the World Bank project-level evaluation data. *World Development 95*, 268–284.

Wane, W. (2004). The quality of foreign aid: Country selectivity or donors incentives? *The World Bank Policy Research Working Paper No. 3325*. https://ssrn.com/abstract=610370.

Winters, M. S. (2019). Too many cooks in the kitchen? The division of financing in World Bank projects and project performance. *Politics and Governance 7*(2), 117–126.

Wood, T., S. Otor, and M. Dornan (2020, 05). Australian aid projects: What works, where projects work and how Australia compares. *Asia & the Pacific Policy Studies 7*, 171–186.

World Bank (2021). World Development Indicators Online Database. https://databank.worldbank.org/source/world-development-indicators.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

A Appendix

A.1 Tables

Table A1: Representativeness of sample

	Non-sample	Sample	Absolute difference
Budget funds (in mil. EUR)	10.15	9.65	0.50
	(10.6)	(11.6)	(1.06)
Share of grants	0.89	0.90	−0.01
	(0.32)	(0.30)	(−0.77)
Disbursement vs. commitment	1.00	1.00	−0.00
	(0.01)	(0.02)	(−0.50)
Time mandate to contract	0.45	0.38	0.06
	(0.86)	(0.83)	(1.79)
Project duration	6.51	6.41	0.10
	(4.14)	(3.89)	(0.58)
Delay	0.19	0.19	0.00
	(0.40)	(0.39)	(0.15)
Observations	1,048	1,124	2,172

Note: The first two columns show the mean values of project characteristics of $N = 1,048$ out-of-sample and $N = 1,124$ in-sample observations. The former represent the sample of projects that were not randomly selected for evaluation in our sample period (cf. section 2). Standard deviations shown in parentheses. Column 3 displays the absolute difference of the mean values in columns one and two. The t-statistic for testing equality of both means is displayed in parentheses below. Significance at or below 1% (***), 5% (**) and 10% (*).

Table A2: Summary statistics

	Full Sample (1)	SSA (2)	Asia/Oceania (3)	Europe/Caucasus (4)	Lat. America (5)	MENA (6)
Financing						
Total volume (in million)	41.66	43.14	58.88	23.57	28.66	29.74
	(125.56)	(146.38)	(159.22)	(51.48)	(59.79)	(44.14)
% counterpart contributions	0.16	0.11	0.18	0.17	0.25	0.15
	(0.21)	(0.17)	(0.23)	(0.22)	(0.21)	(0.18)
Budget Funds (in million)	9.68	7.42	12.96	8.34	7.51	13.91
	(12.10)	(6.30)	(18.53)	(10.08)	(5.88)	(13.12)
% budget funds of ODA (x 1000)	19.79	15.41	21.01	23.81	22.50	24.40
	(29.65)	(18.87)	(38.60)	(33.46)	(27.90)	(31.23)
% project funds of GDP (x 1000)	122.00	157.70	121.09	121.73	51.13	87.22
	(266.94)	(279.71)	(370.38)	(169.39)	(68.22)	(119.43)
Disbursement vs. commitment	0.98	0.99	0.97	0.98	1.00	0.95
	(0.11)	(0.08)	(0.15)	(0.12)	(0.00)	(0.15)
Structure						
Co-financing	0.21	0.30	0.15	0.16	0.20	0.11
	(0.41)	(0.46)	(0.36)	(0.37)	(0.40)	(0.32)
Accompanying measure	0.27	0.20	0.30	0.50	0.17	0.28
	(0.44)	(0.40)	(0.46)	(0.50)	(0.38)	(0.45)
Previous cooperation	0.23	0.29	0.23	0.05	0.17	0.31
	(0.42)	(0.45)	(0.42)	(0.21)	(0.38)	(0.47)
Number of institutions	4.00	4.26	3.67	3.69	4.46	3.79
	(2.52)	(2.66)	(2.02)	(2.37)	(3.01)	(2.52)
Project manager turnover	0.48	0.46	0.49	0.52	0.47	0.51
	(0.38)	(0.38)	(0.30)	(0.47)	(0.35)	(0.33)
Country office	0.46	0.36	0.59	0.47	0.38	0.57
	(0.50)	(0.48)	(0.49)	(0.50)	(0.49)	(0.50)
Complexity						
Project duration	7.04	6.82	7.09	5.69	7.55	8.97
	(3.71)	(3.30)	(3.75)	(3.11)	(4.11)	(4.24)
Delay indicator	0.23	0.19	0.27	0.12	0.21	0.48
	(0.42)	(0.39)	(0.44)	(0.33)	(0.41)	(0.50)
Revised ToC	0.37	0.35	0.37	0.33	0.38	0.44
	(0.48)	(0.48)	(0.48)	(0.47)	(0.49)	(0.50)
Technical complexity	0.48	0.35	0.67	0.63	0.15	0.65
	(0.50)	(0.48)	(0.47)	(0.48)	(0.36)	(0.48)
Risks						
Number ex-ante identified risks	3.99	3.93	4.25	3.67	4.00	4.05
	(2.02)	(1.94)	(2.24)	(1.88)	(2.04)	(1.84)
% ex-ante identified risks occurred	0.55	0.62	0.52	0.49	0.53	0.53
	(0.35)	(0.34)	(0.35)	(0.36)	(0.34)	(0.34)
Macro						
GDP p.c. growth (annual)	3.29	2.47	4.98	4.54	2.37	1.67
	(2.97)	(2.30)	(2.50)	(4.18)	(1.24)	(3.19)
Freedom House Democracy score	4.00	4.05	3.43	4.46	5.30	2.88
	(1.45)	(1.34)	(1.65)	(1.01)	(0.67)	(0.85)
State Fragility Index	12.19	15.00	12.91	7.21	8.46	11.02
	(4.71)	(3.85)	(3.71)	(2.96)	(3.38)	(4.58)
Sectors						
Agr. & Env.	0.13	0.13	0.12	0.07	0.34	0.03
Budget Support	0.02	0.04	0.00	0.00	0.02	0.00
Education	0.07	0.07	0.09	0.02	0.04	0.13
Energy	0.09	0.04	0.14	0.17	0.07	0.08
Finance	0.14	0.07	0.17	0.24	0.15	0.18
Governance	0.08	0.10	0.03	0.09	0.09	0.08
Health	0.13	0.19	0.19	0.03	0.03	0.05
Transportation	0.06	0.08	0.10	0.02	0.02	0.01
Water Supply	0.18	0.17	0.08	0.28	0.14	0.32
Other	0.09	0.11	0.07	0.07	0.10	0.12
Observations	1,124	428	281	167	116	122

Note: Table shows mean and standard deviation in parentheses of covariates, excluding categorial variables and population of the country. % budget funds of ODA and % projects fund of GDP are re-scaled by one thousand. Sector figures correspond to fraction of the respective sector in each sub-sample. Observations are weighted by the inverse of the number of projects evaluated in the corresponding evaluation report. The mean value and standard deviation of the Freedom House Democracy Score and the State Fragility Index is calculated from only 1,096 observations in the full sample due to unavailable data. Cross-regional projects ($N = 10$) are not shown as individual sub-sample but are included in the full sample figures (column 1). Abbreviations: SSA = Sub-Saharan Africa; MENA = Middle East and North Africa.

Table A3: Determinants of success ratings - regional split

Dep. variable: Rating (Pooled)	(1) All	(2) SSA	(3) Asia/Oceania	(4) Europe/Caucasus	(5) Lat. America	(6) MENA
Financing						
Total volume (log)	0.037	0.030	−0.003	0.107	0.074	0.160*
	(0.029)	(0.041)	(0.054)	(0.070)	(0.054)	(0.082)
Aid type (Base: Loan):						
-Grant	0.105	1.151***	−0.157	−0.092	0.359**	−0.536**
	(0.087)	(0.271)	(0.136)	(0.153)	(0.147)	(0.215)
% counterpart contributions	0.145	0.005	0.232	0.322	0.457	1.665***
	(0.118)	(0.191)	(0.238)	(0.345)	(0.430)	(0.256)
Budget funds (log)	0.095**	0.047	0.140**	0.179**	−0.025	0.210*
	(0.042)	(0.063)	(0.069)	(0.086)	(0.085)	(0.122)
% budget funds of ODA	−0.000	−0.000	−0.000	−0.000***	0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
% project funds of GDP	−0.000	0.000	−0.000	−0.000*	−0.000	−0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Disbursement vs. commitment	0.137	−0.398	0.457	0.048	–	–
	(0.156)	(0.338)	(0.309)	(0.394)	(–)	(–)
Structure						
Co-financing	0.002	−0.006	−0.104	0.298**	−0.183	−0.365***
	(0.064)	(0.082)	(0.137)	(0.133)	(0.200)	(0.098)
Accompanying measure	−0.015	0.105	−0.123	−0.139*	−0.016	−0.502***
	(0.056)	(0.090)	(0.107)	(0.071)	(0.165)	(0.182)
Agency type (Base: NGO):						
-Mixed	−0.099	−0.324*	0.074	0.325	−0.073	−0.043
	(0.130)	(0.173)	(0.253)	(0.281)	(0.199)	(0.593)
-Multilateral	−0.009	−0.184	0.776	0.196	0.133	0.362
	(0.131)	(0.200)	(0.520)	(0.360)	(0.295)	(0.255)
-Private sector	0.006	−0.282	0.467	−0.247	0.139	2.156***
	(0.139)	(0.213)	(0.329)	(0.286)	(0.344)	(0.416)
-Government	−0.101	−0.313**	0.277	−0.226	−0.162	−0.209
	(0.107)	(0.142)	(0.219)	(0.259)	(0.182)	(0.464)
Previous cooperation	0.066	−0.011	0.050	0.399**	0.107	0.279**
	(0.051)	(0.071)	(0.108)	(0.195)	(0.125)	(0.116)
Number of institutions	0.005	0.011	−0.011	−0.050**	−0.031*	0.019
	(0.009)	(0.015)	(0.025)	(0.023)	(0.017)	(0.031)
Project manager turnover	0.328	0.702	0.694	−0.590***	0.511	−0.770***
	(0.248)	(0.533)	(0.449)	(0.181)	(0.809)	(0.216)
Country office	−0.043	−0.070	−0.047	0.003	0.234*	0.490**
	(0.056)	(0.111)	(0.122)	(0.118)	(0.119)	(0.194)
Complexity						
Project duration (log)	−0.149**	0.124	−0.037	−0.372**	−0.564***	−0.143
	(0.075)	(0.150)	(0.135)	(0.148)	(0.149)	(0.119)
Delay indicator	0.009	0.001	−0.149	0.167	0.466**	−0.069
	(0.069)	(0.104)	(0.125)	(0.216)	(0.228)	(0.118)
Revised ToC	−0.048	−0.034	−0.035	−0.265***	−0.003	0.447***
	(0.047)	(0.079)	(0.095)	(0.086)	(0.113)	(0.131)
Years mandate to contract	−0.048*	−0.065	0.027	−0.109*	−0.038	0.048
	(0.027)	(0.070)	(0.037)	(0.060)	(0.034)	(0.097)
Technical complexity	−0.130**	−0.227***	0.042	0.090	−0.267*	0.131
	(0.055)	(0.076)	(0.121)	(0.161)	(0.151)	(0.177)
Risks						
Number ex-ante identified risks	0.001	−0.025	−0.013	−0.024	0.040	−0.040
	(0.013)	(0.021)	(0.025)	(0.029)	(0.026)	(0.026)
% ex-ante identified risks occurred	−0.486***	−0.525***	−0.478***	−0.264**	−0.839***	−0.412**
	(0.067)	(0.119)	(0.135)	(0.122)	(0.182)	(0.159)
Overall risk (base: low)						
-Medium	−0.203**	−0.523***	0.122	−0.534**	0.480	−0.202
	(0.082)	(0.145)	(0.172)	(0.212)	(0.297)	(0.396)
-(Very) high	−0.352***	−0.655***	−0.131	−0.572**	0.436	−0.197
	(0.088)	(0.149)	(0.199)	(0.220)	(0.305)	(0.410)
-Not assigned	−0.219*	−0.305*	−0.173	−0.245	0.120	−0.479
	(0.116)	(0.174)	(0.208)	(0.257)	(0.341)	(0.557)
Overall risk control (base: low)						
-Medium	0.084	0.125	0.212*	0.167	−0.129	0.048
	(0.058)	(0.095)	(0.112)	(0.144)	(0.139)	(0.247)
-High	−0.061	0.058	−0.657***	0.760***		−0.189
	(0.169)	(0.230)	(0.228)	(0.274)		(0.744)
Macro variables						
GDP p.c. growth (annual)	0.011	0.016	0.042*	0.010	0.146***	0.025
	(0.008)	(0.015)	(0.022)	(0.012)	(0.038)	(0.022)
Freedom House Democracy score	−0.018	0.001	−0.019	−0.208***	−0.490***	−0.036
	(0.021)	(0.049)	(0.037)	(0.064)	(0.120)	(0.066)
State Fragility Index	−0.006	−0.017	0.048***	0.029	−0.072**	0.002
	(0.008)	(0.018)	(0.018)	(0.029)	(0.032)	(0.023)
Population log	−0.029	0.050	−0.074**	−0.043	−0.002	−0.281**
	(0.022)	(0.059)	(0.036)	(0.062)	(0.052)	(0.134)
Sector indicators	Yes	Yes	Yes	Yes	Yes	Yes
Sub-rating indicators	Yes	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,458	2,136	1,401	804	580	487
Adjusted R^2	0.23	0.29	0.27	0.51	0.54	0.57

Note: Table entries are coefficients from WLS regressions with the pooled rating as dependent variable. Weights are given by the inverse of the number of projects evaluated in the corresponding evaluation report. Observations are weighted by the inverse of the number of projects evaluated in the corresponding evaluation report. Other control variables include: Number of years between final project inspection and evaluation; the year of project start as well as evaluation year (both 5-year intervals); evaluation month. Standard errors in parentheses clustered at the country-evaluation-year-level. Significance at or below 1% (***), 5% (**) and 10% (*).

Table A4: Determinants of success ratings - sectoral split (reduced)

Dep. variable: Rating (Pooled)	(1) Full Sample	(2) Agr. & Env.	(3) Education	(4) Energy	(5) Finance	(6) Health	(7) Governance	(8) Transportation	(9) Water Supply	(10) Other
Financing										
Total volume (log)	0.016	0.084	−0.035	0.220***	−0.060	−0.020	−0.150***	−0.022	0.056	0.149**
	(0.029)	(0.092)	(0.061)	(0.046)	(0.068)	(0.054)	(0.051)	(0.072)	(0.089)	(0.063)
% counterpart contributions	0.191	1.175***	0.177	0.500**	−0.099	−0.087	0.445*	0.493	0.596*	−0.814***
	(0.117)	(0.433)	(0.400)	(0.234)	(0.402)	(0.210)	(0.264)	(0.341)	(0.346)	(0.255)
Budget funds (log)	0.115***	0.045	0.066	0.287***	0.064	−0.035	0.228**	−0.562***	0.159	−0.079
	(0.041)	(0.114)	(0.158)	(0.079)	(0.089)	(0.118)	(0.087)	(0.147)	(0.140)	(0.081)
% budget funds of ODA	−0.000	−0.000**	−0.000	−0.000***	0.000	0.000	−0.000	0.000**	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
% project funds of GDP	−0.000	0.000	0.000***	0.000	−0.000	0.000	0.000	0.000***	−0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Structure										
Co-financing	0.024	−0.136	−0.285*	0.155	0.005	−0.031	−0.030	−0.236	0.161	0.054
	(0.062)	(0.178)	(0.158)	(0.116)	(0.213)	(0.130)	(0.230)	(0.166)	(0.208)	(0.142)
Accompanying measure	−0.028	−0.001	0.315	0.063	0.099	−0.211	−0.408***	−0.063	−0.030	−0.800***
	(0.056)	(0.142)	(0.195)	(0.100)	(0.144)	(0.182)	(0.140)	(0.260)	(0.112)	(0.143)
Number of institutions	0.002	−0.051*	−0.009	−0.119***	−0.029	−0.007	0.009	0.251***	0.014	−0.007
	(0.009)	(0.026)	(0.022)	(0.028)	(0.033)	(0.022)	(0.017)	(0.045)	(0.027)	(0.021)
Project manager turnover	0.368	2.247*	−2.775***	0.195	−0.137	0.152	−0.499	0.810	−0.244	−0.107
	(0.264)	(1.165)	(0.757)	(0.445)	(0.352)	(0.790)	(0.528)	(1.361)	(0.496)	(0.608)
Country office	−0.054	0.025	−0.197	−0.490***	−0.036	−0.081	−0.025	0.839***	−0.144	−0.152
	(0.056)	(0.168)	(0.260)	(0.156)	(0.159)	(0.112)	(0.126)	(0.255)	(0.150)	(0.096)
Complexity										
Project duration (log)	−0.138*	0.241	−0.100	−0.344*	−0.463***	0.362**	0.061	−0.007	−0.466**	−0.049
	(0.076)	(0.307)	(0.249)	(0.191)	(0.150)	(0.155)	(0.141)	(0.277)	(0.233)	(0.137)
Delay indicator	−0.190***	−0.131	−0.618***	0.160	−0.287	−0.285*	0.042	−0.492	0.158	−0.339**
	(0.065)	(0.219)	(0.224)	(0.148)	(0.226)	(0.167)	(0.122)	(0.306)	(0.172)	(0.146)
Revised ToC	−0.066	0.095	−0.229*	−0.026	−0.043	−0.068	0.042	−0.252	−0.235**	−0.197*
	(0.047)	(0.134)	(0.128)	(0.122)	(0.192)	(0.087)	(0.144)	(0.170)	(0.106)	(0.107)
Years mandate to contract	−0.050*	0.064	0.477***	0.122***	−0.065	−0.063	−0.336***	0.080	−0.168**	0.153**
	(0.026)	(0.078)	(0.148)	(0.041)	(0.067)	(0.080)	(0.069)	(0.125)	(0.069)	(0.075)
Technical complexity	−0.123**	−0.361	0.046	−0.226	0.323	−0.112	−0.221*	−0.021	−0.145	−0.480***
	(0.054)	(0.223)	(0.201)	(0.155)	(0.284)	(0.132)	(0.129)	(0.229)	(0.185)	(0.094)
Risks										
Number ex-ante identified risks	−0.002	0.030	−0.127***	−0.037**	−0.003	−0.027	0.014	−0.031	0.004	0.065**
	(0.013)	(0.041)	(0.043)	(0.019)	(0.041)	(0.030)	(0.030)	(0.046)	(0.028)	(0.031)
% ex-ante identified risks occured	−0.503***	−0.669***	−0.841***	0.337*	−0.327	−0.549***	−0.473**	−0.670***	−0.735***	−0.362**
	(0.070)	(0.202)	(0.173)	(0.186)	(0.205)	(0.157)	(0.214)	(0.150)	(0.149)	(0.168)
Macro variables										
GDP p.c. growth (annual)	0.012	0.023	0.010	−0.033**	0.005	0.009	0.012	0.001	0.013	−0.034
	(0.008)	(0.032)	(0.021)	(0.016)	(0.014)	(0.027)	(0.014)	(0.024)	(0.028)	(0.031)
Freedom House Democracy score	−0.020	0.006	−0.054	−0.067*	−0.042	−0.052	0.091	−0.245**	−0.056	0.020
	(0.022)	(0.059)	(0.087)	(0.038)	(0.058)	(0.042)	(0.092)	(0.092)	(0.065)	(0.044)
State Fragility Index	−0.008	−0.065**	−0.078**	−0.013	0.027	−0.029*	0.016	−0.043	−0.000	−0.017
	(0.008)	(0.025)	(0.034)	(0.020)	(0.026)	(0.015)	(0.022)	(0.039)	(0.018)	(0.019)
Population (log)	−0.029	0.061	0.125**	−0.208***	−0.039	0.057	−0.077	0.264***	−0.021	−0.043
	(0.022)	(0.064)	(0.054)	(0.040)	(0.062)	(0.070)	(0.062)	(0.082)	(0.062)	(0.044)
Region indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sub-rating indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,283	654	370	455	815	803	503	340	824	519
Adjusted R^2	0.22	0.38	0.46	0.50	0.30	0.39	0.54	0.59	0.37	0.53

Note: Table entries are coefficients from WLS regressions with the pooled rating as dependent variable. Weights are given by the inverse of the number of projects evaluated in the corresponding evaluation report. Reduced set of covariates (exclusion of categorical variables, 'Disbursement vs. commitment' and 'Previous cooperation') due to lack of variation in small sub-samples. Sector 'Budget support' is not displayed as sub-sample for the same reason. Other control variables include: Number of years between final project inspection and evaluation; the year of project start as well as evaluation year (both 5-year intervals); evaluation month. Standard errors in parentheses are clustered at the country-evaluation-year-level. Significance at or below 1% (***), 5% (**) and 10% (*).

Table A5: OECD DAC Ratings

	(1) Overall	(2) Relevance	(3) Efficiency	(4) Effectiveness	(5) Impact	(6) Sustainability
Financing						
Total volume (log)	0.037	0.015	0.047	0.053	0.014	0.055*
	(0.029)	(0.034)	(0.041)	(0.043)	(0.036)	(0.032)
Aid type (Base: Loan):						
-Grant	0.105	0.034	0.147	0.155	0.146	0.039
	(0.087)	(0.102)	(0.126)	(0.126)	(0.115)	(0.106)
% counterpart contributions	0.145	0.156	−0.019	0.283*	0.213	0.100
	(0.118)	(0.152)	(0.166)	(0.163)	(0.171)	(0.148)
Budget funds (log)	0.095**	0.054	0.092*	0.133**	0.127***	0.067
	(0.042)	(0.052)	(0.054)	(0.059)	(0.048)	(0.048)
% budget funds of ODA	−0.000	0.000	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
% project funds of GDP	−0.000	−0.000	−0.000	−0.000	−0.000	−0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Disbursement vs. commitment	0.137	0.019	0.212	0.207	0.184	0.048
	(0.156)	(0.178)	(0.250)	(0.267)	(0.228)	(0.261)
Structure						
Cofinancing	0.002	−0.064	−0.009	0.028	0.076	−0.013
	(0.064)	(0.081)	(0.082)	(0.088)	(0.084)	(0.072)
Accompanying measure	−0.015	0.032	0.001	−0.062	−0.007	−0.040
	(0.056)	(0.062)	(0.078)	(0.079)	(0.079)	(0.068)
Agency type (Base: NGO):						
-Mixed	−0.099	−0.038	−0.168	−0.103	−0.078	−0.106
	(0.130)	(0.142)	(0.191)	(0.181)	(0.169)	(0.138)
-Multilateral	−0.009	0.123	−0.074	−0.191	0.040	0.051
	(0.131)	(0.156)	(0.206)	(0.175)	(0.177)	(0.153)
-Private sector	0.006	−0.116	0.023	0.048	0.062	0.010
	(0.139)	(0.159)	(0.202)	(0.185)	(0.186)	(0.147)
-Government	−0.101	−0.079	−0.128	−0.055	−0.097	−0.144
	(0.107)	(0.119)	(0.160)	(0.150)	(0.146)	(0.106)
Previous cooperation	0.066	−0.025	0.131*	0.027	0.086	0.115**
	(0.051)	(0.059)	(0.074)	(0.074)	(0.069)	(0.058)
Number of institutions	0.005	0.017	0.003	−0.003	0.019	−0.012
	(0.009)	(0.011)	(0.013)	(0.011)	(0.012)	(0.010)
Project manager turnover	0.328	0.237	0.258	0.640	0.631*	−0.129
	(0.248)	(0.277)	(0.278)	(0.395)	(0.322)	(0.252)
Country office	−0.043	−0.000	0.024	−0.128*	−0.048	−0.059
	(0.056)	(0.067)	(0.076)	(0.076)	(0.076)	(0.065)
Complexity						
Project duration (log)	−0.149**	−0.077	−0.150	−0.178*	−0.032	−0.314***
	(0.075)	(0.084)	(0.107)	(0.101)	(0.097)	(0.082)
Delay indicator	0.009	0.121	−0.124	−0.086	0.121	0.018
	(0.069)	(0.082)	(0.093)	(0.092)	(0.089)	(0.081)
Revised ToC	−0.048	−0.106*	−0.018	−0.028	−0.110*	0.026
	(0.047)	(0.058)	(0.066)	(0.064)	(0.066)	(0.054)
Years mandate to contract	−0.048*	−0.038	−0.034	−0.082***	−0.058	−0.028
	(0.027)	(0.031)	(0.038)	(0.031)	(0.038)	(0.033)
Technical complexity	−0.130**	−0.013	−0.215***	−0.109	−0.215***	−0.099
	(0.055)	(0.064)	(0.081)	(0.078)	(0.078)	(0.063)
Risks						
Number ex-ante identified risks	0.001	0.001	−0.002	−0.002	0.014	−0.004
	(0.013)	(0.014)	(0.018)	(0.018)	(0.016)	(0.014)
% ex-ante identified risks occurred	−0.486***	−0.264***	−0.593***	−0.558***	−0.480***	−0.537***
	(0.067)	(0.081)	(0.092)	(0.097)	(0.094)	(0.075)
Overall risk (base:low)						
-Medium	−0.203**	−0.011	−0.326**	−0.131	−0.260**	−0.284**
	(0.082)	(0.098)	(0.139)	(0.111)	(0.121)	(0.111)
-(Very) high	−0.352***	−0.080	−0.494***	−0.344***	−0.435***	−0.407***
	(0.088)	(0.112)	(0.146)	(0.120)	(0.133)	(0.117)
-Not assigned	−0.219*	0.017	−0.349*	−0.124	−0.259	−0.384**
	(0.116)	(0.149)	(0.178)	(0.162)	(0.174)	(0.151)
Overall risk control (base: low)						
-Medium	0.084	0.121*	0.050	0.103	0.160*	−0.022
	(0.058)	(0.070)	(0.082)	(0.081)	(0.083)	(0.073)
-High	−0.061	0.140	−0.300	−0.085	0.001	−0.060
	(0.169)	(0.219)	(0.239)	(0.169)	(0.226)	(0.211)
Macro variables						
GDP p.c. growth (annual)	0.011	0.003	0.020*	0.002	0.014	0.018**
	(0.008)	(0.008)	(0.011)	(0.011)	(0.011)	(0.009)
Freedom House Democracy score	−0.018	0.029	−0.034	−0.039	−0.030	−0.016
	(0.021)	(0.024)	(0.031)	(0.028)	(0.031)	(0.025)
State Fragility Index	−0.006	0.016	−0.007	−0.006	−0.009	−0.020**
	(0.008)	(0.010)	(0.011)	(0.011)	(0.011)	(0.009)
Population log	−0.029	−0.029	−0.047	−0.030	−0.037	−0.003
	(0.022)	(0.026)	(0.029)	(0.028)	(0.031)	(0.025)
Sub-rating indicators	Yes					
Sector and region indicators	Yes	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,458	1,092	1,094	1,093	1,093	1,086
Adjusted R^2	0.23	0.07	0.15	0.16	0.16	0.21

Note: Table entries are coefficients from WLS regressions with individual DAC criteria as dependent variable. Weights are given by the inverse of the number of projects evaluated in the corresponding evaluation report. Other control variables include: Number of years between final project inspection and evaluation; the year of project start as well as evaluation year (both 5-year intervals); evaluation month. Standard errors in parentheses are clustered at the country-evaluation-year-level. Significance at or below 1% (***), 5% (**) and 10% (*).

Table A6: Robustness - Main outcome variables as OLS and ordered probit models

	Rating (pooled)			Overall Rating		Arithmetic Rating		Binary
	(1) OLS	(2) O. Probit	(3) Vol. weighted	(4) OLS	(5) O. Probit	(6) OLS	(7) O. Probit	(8) Probit
Financing								
Total volume (log)	0.037	0.020	0.012	0.050	0.028	0.037	0.012	0.061
	(0.029)	(0.034)	(0.022)	(0.041)	(0.048)	(0.030)	(0.042)	(0.080)
Aid type (Base: Loan):								
-Grant	0.105	0.073	0.042	0.129	0.095	0.104	0.080	0.055
	(0.087)	(0.132)	(0.087)	(0.121)	(0.165)	(0.089)	(0.163)	(0.239)
% counterpart contributions	0.145	0.050	0.026	0.176	0.073	0.149	0.047	0.530
	(0.118)	(0.170)	(0.109)	(0.170)	(0.235)	(0.122)	(0.215)	(0.355)
Budget funds (log)	0.095**	0.128**	0.085**	0.123**	0.176**	0.096**	0.157**	0.355***
	(0.042)	(0.057)	(0.037)	(0.059)	(0.074)	(0.043)	(0.071)	(0.107)
% budget funds of ODA	−0.000	−0.000	−0.000	−0.000*	−0.000*	−0.000	−0.000	−0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
% project funds of GDP	−0.000	−0.000	−0.000	−0.000	−0.000	−0.000	−0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Disbursement vs. commitment	0.137	0.282	0.228	0.286	0.386	0.128	0.365	1.112**
	(0.156)	(0.302)	(0.212)	(0.234)	(0.366)	(0.159)	(0.377)	(0.501)
Structure								
Co-financing	0.002	0.044	0.028	0.083	0.221**	−0.001	0.059	0.274
	(0.064)	(0.081)	(0.053)	(0.088)	(0.112)	(0.066)	(0.101)	(0.171)
Accompanying measure	−0.015	−0.048	−0.037	−0.078	−0.109	−0.013	−0.053	−0.228
	(0.056)	(0.074)	(0.050)	(0.077)	(0.104)	(0.057)	(0.090)	(0.151)
Agency type (Base: NGO):								
-Mixed	−0.099	−0.070	−0.046	−0.092	−0.089	−0.103	−0.065	−0.384
	(0.130)	(0.177)	(0.118)	(0.180)	(0.241)	(0.134)	(0.223)	(0.353)
-Multilateral	−0.009	0.249	0.137	−0.045	0.326	−0.011	0.297	0.586
	(0.131)	(0.220)	(0.136)	(0.186)	(0.300)	(0.134)	(0.276)	(0.588)
-Private sector	0.006	−0.067	−0.049	0.049	−0.014	0.004	−0.032	−0.371
	(0.139)	(0.219)	(0.148)	(0.185)	(0.279)	(0.142)	(0.264)	(0.377)
-Government	−0.101	−0.072	−0.048	−0.094	−0.075	−0.101	−0.059	−0.389
	(0.107)	(0.149)	(0.099)	(0.145)	(0.199)	(0.110)	(0.185)	(0.298)
Previous cooperation	0.066	0.066	0.050	0.112	0.125	0.065	0.074	0.163
	(0.051)	(0.068)	(0.045)	(0.070)	(0.092)	(0.052)	(0.084)	(0.143)
Number of institutions	0.005	0.021	0.013	0.006	0.020	0.005	0.029*	0.013
	(0.009)	(0.013)	(0.009)	(0.011)	(0.016)	(0.009)	(0.016)	(0.026)
Project manager turnover	0.328	0.426	0.291	0.449*	0.548*	0.321	0.397	2.279**
	(0.248)	(0.328)	(0.215)	(0.267)	(0.319)	(0.256)	(0.382)	(0.899)
Country office	−0.043	−0.087	−0.058	−0.093	−0.180*	−0.044	−0.115	−0.400***
	(0.056)	(0.080)	(0.053)	(0.076)	(0.104)	(0.057)	(0.100)	(0.149)
Complexity								
Project duration (log)	−0.149**	−0.230**	−0.148**	−0.202**	−0.310**	−0.146*	−0.323***	−0.387*
	(0.075)	(0.099)	(0.063)	(0.099)	(0.130)	(0.078)	(0.123)	(0.200)
Delay indicator	0.009	−0.013	−0.022	−0.010	−0.036	0.005	−0.005	0.012
	(0.069)	(0.104)	(0.069)	(0.094)	(0.129)	(0.071)	(0.125)	(0.174)
Revised ToC	−0.048	−0.081	−0.050	−0.078	−0.095	−0.049	−0.101	−0.066
	(0.047)	(0.070)	(0.047)	(0.064)	(0.094)	(0.048)	(0.087)	(0.141)
Years mandate to contract	−0.048*	−0.048	−0.031	−0.046	−0.049	−0.049*	−0.064	−0.057
	(0.027)	(0.037)	(0.025)	(0.037)	(0.047)	(0.028)	(0.046)	(0.066)
Technical complexity	−0.130**	−0.214***	−0.140***	−0.121	−0.246**	−0.130**	−0.240**	−0.454***
	(0.055)	(0.078)	(0.052)	(0.076)	(0.103)	(0.057)	(0.094)	(0.157)
Risks								
Number ex-ante identified risks	0.001	0.006	0.003	0.000	0.002	0.001	0.008	−0.034
	(0.013)	(0.019)	(0.012)	(0.017)	(0.024)	(0.013)	(0.024)	(0.032)
% ex-ante identified risks occurred	−0.486***	−0.760***	−0.504***	−0.650***	−1.008***	−0.486***	−0.940***	−1.548***
	(0.067)	(0.103)	(0.069)	(0.092)	(0.136)	(0.069)	(0.128)	(0.209)
Overall risk (base: low)								
-Medium	−0.203**	−0.292**	−0.187**	−0.292**	−0.457**	−0.203**	−0.354**	−1.237***
	(0.082)	(0.128)	(0.080)	(0.124)	(0.201)	(0.084)	(0.166)	(0.440)
-(Very) high	−0.352***	−0.519***	−0.333***	−0.496***	−0.767***	−0.349***	−0.629***	−1.459***
	(0.088)	(0.134)	(0.084)	(0.131)	(0.211)	(0.091)	(0.174)	(0.465)
- Not assigned	−0.219*	−0.424***	−0.276***	−0.349**	−0.630**	−0.217*	−0.550***	−1.312**
	(0.116)	(0.163)	(0.103)	(0.175)	(0.249)	(0.119)	(0.211)	(0.511)
Overall risk control (base: low)								
-Medium	0.084	0.125	0.077	0.140*	0.182	0.086	0.143	0.101
	(0.058)	(0.085)	(0.055)	(0.076)	(0.112)	(0.060)	(0.107)	(0.155)
-High	−0.061	0.074	0.034	−0.106	−0.005	−0.061	0.074	−0.070
	(0.169)	(0.284)	(0.175)	(0.221)	(0.355)	(0.173)	(0.367)	(0.435)
Macro variables								
GDP p.c. growth (annual)	0.011	0.012	0.009	0.009	0.012	0.011	0.012	0.037
	(0.008)	(0.012)	(0.008)	(0.011)	(0.016)	(0.008)	(0.015)	(0.023)
Freedom House Democracy score	−0.018	−0.034	−0.022	−0.031	−0.068	−0.019	−0.052	−0.056
	(0.021)	(0.036)	(0.024)	(0.029)	(0.047)	(0.022)	(0.044)	(0.059)
State Fragility Index	−0.006	−0.006	−0.005	−0.009	−0.014	−0.006	−0.011	−0.028
	(0.008)	(0.013)	(0.009)	(0.010)	(0.017)	(0.008)	(0.016)	(0.022)
Population (log)	−0.029	−0.036	−0.024	−0.031	−0.042	−0.029	−0.036	−0.085
	(0.022)	(0.039)	(0.025)	(0.028)	(0.048)	(0.022)	(0.048)	(0.058)
Sector and region indicators	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sub-rating indicators	Yes	Yes	Yes					
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5,458	5,458	5,458	1,094	1,094	1,094	1,094	1,053
Adjusted R^2	0.23		0.24	0.19		0.19		

Note: Table entries are coefficients from WLS regressions with individual DAC criteria as dependent variable. Weights are given by the inverse of the number of projects evaluated in the corresponding evaluation report. Other control variables include: number of years between final project inspection and evaluation; the year of project start as well as evaluation year (both 5-year intervals); evaluation month. Standard errors in parentheses are clustered at the country-evaluation-year-level. Significance at or below 1% (***), 5% (**) and 10% (*).

Table A7: Determinants of success ratings - Lasso Estimates

	(1) Overall	(2) Lasso estimation	(3) Reduced-form estimation
Financing			
Total volume (log)	0.037	0.042	0.047*
	(0.029)		(0.025)
Aid type (base: Loan):			
-Loan		−0.115	−0.127
			(0.084)
-Grant	0.105		
	(0.087)		
% counterpart contributions	0.145		
	(0.118)		
Budget funds (log)	0.095**	0.075	0.083**
	(0.042)		(0.039)
% budget funds of ODA	−0.000	−0.000	−0.000
	(0.000)		(0.000)
% project funds of GDP	−0.000	−0.000	−0.000
	(0.000)		(0.000)
Disbursement vs. commitment	0.137	0.062	0.131
	(0.156)		(0.156)
Structure			
Co-financing	0.002		
	(0.064)		
Accompanying measure	−0.015		
	(0.056)		
Agency type (base: NGO):			
-NGO		0.073	0.095
			(0.106)
-Mixed	−0.099	−0.001	−0.013
	(0.130)		(0.081)
-Multilateral	−0.009		
	(0.131)		
-Private sector	0.006	0.084	0.100
	(0.139)		(0.095)
-Government	−0.101		
	(0.107)		
Previous cooperation	0.066	0.058	0.064
	(0.051)		(0.051)
Number of institutions	0.005	0.002	0.005
	(0.009)		(0.009)
Project manager turnover	0.328	0.161	0.163**
	(0.248)		(0.066)
Country office	−0.043	−0.031	−0.050
	(0.056)		(0.054)
Complexity			
Project duration (log)	−0.149**	−0.154	−0.160**
	(0.075)		(0.073)
Delay indicator	0.009		
	(0.069)		
Revised ToC	−0.048	−0.036	−0.043
	(0.047)		(0.047)
Years mandate to contract	−0.048*	−0.043	−0.046*
	(0.027)		(0.027)
Technical complexity	−0.130**	−0.135	−0.139**
	(0.055)		(0.055)
Risks			
Number ex-ante identified risks	0.001		
	(0.013)		
% ex-ante identified risks occurred	−0.486***	−0.493	−0.492***
	(0.067)		(0.068)
Overall risk (base: low):			
-Medium	−0.203**	−0.172	−0.204**
	(0.082)		(0.079)
-(Very) high	−0.352***	−0.312	−0.344***
	(0.088)		(0.084)
-Not assigned	−0.219*	−0.161	−0.206*
	(0.116)		(0.114)
Overall risk control (base: low):			
-Medium	0.084	0.091	0.089
	(0.058)		(0.056)
-High	−0.061		
	(0.169)		
Macro variables			
GDP p.c. growth (annual)	0.011	0.011	0.011
	(0.008)		(0.008)
Freedom House Democracy score	−0.018	−0.013	−0.021
	(0.021)		(0.020)
State Fragility Index	−0.006	−0.005	−0.007
	(0.008)		(0.008)
Population log	−0.029	−0.019	−0.022
	(0.022)		(0.021)
Sector indicators	Yes	Yes	Yes
Sub-rating indicators	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes
Observations	5,458	5,458	5,458
Adjusted R^2	0.23		0.19

Note: Table entries are coefficients from WLS regressions with the pooled rating as dependent variable (column 1). LASSO (column 2) presents results from an adaptive LASSO regression. Reduced form estimates (column 3) runs the WLS regression on all variables with coefficients that are different from zero in the LASSO regression. Weights are given by the inverse of the number of projects evaluated in the corresponding evaluation report. Other control variables include: Number of years between final project inspection and evaluation; the year of project start as well as evaluation year (both 5 year intervals); evalua-

Table A8: Codebook: Outcome and project variables

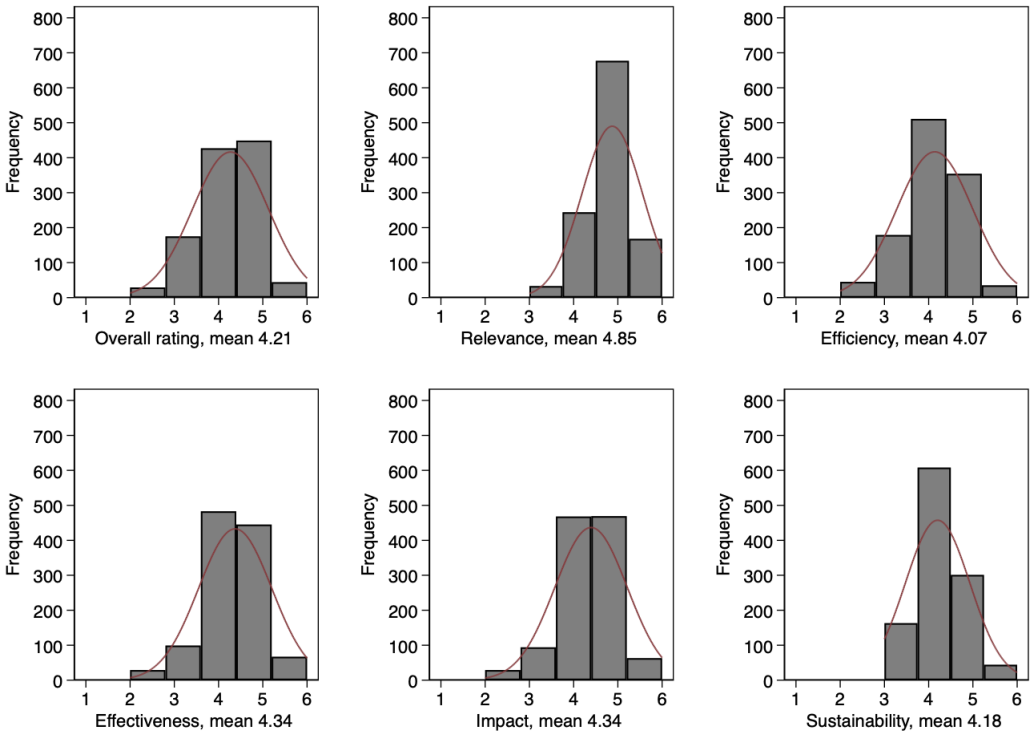
Label	Description	Source
Outcome variables		
Rating (pooled)	Discrete variable on a scale from 1 to 6. Each of the five DAC ratings of each project is considered as an individual observation	Evaluation reports.
Rating	Reported overall ratings. Differs from mean for following reasons: • KO-criterion: For a project to receive a rating of 4 or better, the sub-ratings 'sustainability', 'impact' and 'effectiveness' must not be rated worse than 4 • Differing weights for certain sub-ratings for most projects	Evaluation reports.
Rating (mean)	Continuous variable ranging 1 through 6. Mean of reported DAC ratings.	Evaluation reports.
Relevance	Discrete variable on a scale from 6 to 1.	Evaluation reports.
Effectiveness	Discrete variable on a scale from 6 to 1.	Evaluation reports.
Efficiency	Discrete variable on a scale from 6 to 1.	Evaluation reports.
Impact	Discrete variable on a scale from 6 to 1.	Evaluation reports.
Sustainability	Discrete variable on a scale from 6 to 1.	Evaluation reports.
Micro variables		
Aid type	Categorical variable. Aid types: (1) Loan, (2) Grant	Evaluation reports.
Total volume (log)	Discrete variable. Sum of project volume and counterpart contribution. In current EUR million, logarithmized.	Evaluation reports.
Budget funds (log)	Discrete variable. Budget funds in current EUR million, logged.	KfW data.
Share of counterpart contribution	Discrete variable. Share of counterpart contribution relative to investment volume.	KfW data.
% budget Funds of ODA	Continuous variable. Share of budget funds relative to (current) EUR value of ODA commitments (grants) in the year of project start in a given country (multiplied by 1 million).	KfW data and OECD Statistics (OECD, 2022a).
% project funds of GDP	Continuous variable. Share of project volume relative to the country's GDP in the year of project start (multiplied by 1 billion). If the project operates in more than one country, the sum of GDP across countries is taken as reference value.	KfW data and World Bank national accounts data, and OECD National Accounts data files (World Bank, 2021).
Disbursement vs. commitment	Continuous variable. Share of disbursements relative to commitments.	KfW data.
Co-financing	Binary variable. Indicates co-financing: 0 = no. 1 = yes.	Evaluation reports.
Accompanying measure	Binary variable. Indicates accompanying measure. 0 = no. 1 = yes.	Evaluation reports.
Agency type	Categorical variable. Type of programme executing agency, gov.,multilateral, NGO, private or mixed.	Evaluation reports.
Number of institutions	Discrete (integer) variable. Number of institutions involved.	Evaluation reports.
Previous cooperation	Binary variable. Previous cooperation with implementing agency.	KfW data.
Project manager turnover	Continuous variable. Number of project managers assigned to project divided by project duration.	KfW data.
Country office	Binary variable for presence of KfW office in project country during entire project implementation	KfW data.
Number of ex ante identified risks	Discrete variable. Count of ex-ante identified risks.	Evaluation reports.
% ex ante identified risks that occurred	Continuous variable. Share of ex-ante identified risks that materialised.	Evaluation reports.
Overall risk	Categorical variable. Ex-ante overall project risk assessment by project manager	KfW data.
Overall risk control	Categorical variable. Ex-ante overall project risk assessment of controllability by project manager	KfW data.
Project duration (log)	Discrete variable. Year of first contract signed until year of final review, logarithmised.	KfW data.
Delay Indicator	Binary variable. 'Delay' is identified as follows: • Identification of 80th percentile of project duration for each sector in 10-year intervals. • Projects are classified 'delayed' if project duration was longer than the 80th percentile of project durations in the sector. Sectoral comparison group needs to meet the following criteria: • Projects with project start year >= 1990. • Finished projects (existing final review except for financial sector).	Evaluation reports.
Revised ToC	Binary variable. This variable indicates whether the Theory of Change was revised, characterized by either or both of the following: • Impact objective of the project has changed. • Outcome objective of the project has changed.	Evaluation reports.
Years mandate to contract	Discrete variable. Time between mandate and first contract in years.	KfW data.
Technical complexity	Binary variable. Indicates involvement of technical expert: 0 = no. 1 = yes.	Evaluation reports.

Table A9: Codebook: Macro, control and analytical variables

Label	Description	Source
Macro variables		
Population	Logarithm of population.	World Development Indicators (World Bank, 2021).
GDP p.c. growth (annual)	GDP per capita growth (annual %).	World Bank national accounts data, and OECD National Accounts data files (World Bank, 2021).
Democracy	Mean of score for: • Political rights • Civil liberties Interpretation: • Value 1 – 2.5: Free • Value 3 – 5.5: Partly free • Value 5.6 – 7: Not free	Freedom House: Freedom in the World (Freedom House, 2021)
Fragility Index	State Fragility Index as sum of Effectiveness Score + Legitimacy Score (25 points possible) • Effectiveness Score = Security Effectiveness + Political Effectiveness + Economic Effectiveness + Social Effectiveness (13 points possible) • Legitimacy Score = Security Legitimacy + Political Legitimacy + Economic Legitimacy + Social Legitimacy (12 points possible) Interpretation: higher value corresponds to higher fragility • Value 20–25: Extreme fragility • Value 16–19: High fragility • Value 12–15: Serious • Value 8–11: Moderate • Value 4–7: Low • Value 0–3: Little or no	Center for Systemic Peace (Integrated Network for Societal Conflict Research (INSCR), 2018).
Net ODA (commitments: Grants)	Net Official Development Aid (ODA) commitments (grants only) in the year of project start. If a project operated in more than one country, the value of this variable corresponds to the sum of ODA in respective countries.	OECD Statistics (OECD, 2022a).
Evaluation and Control variables		
Sector	Projects are either assigned one of nine sectors or classified as 'other' if none of the sectors are applicable. List of sectors: Budget Support, education, energy, finance, health & population, governance, agriculture & environment, transportation, Water Supply.	Evaluation reports.
Region	Region of the project. List of regions: Sub-Saharan Africa, Asia/Oceania, Europe/Caucasus, Latin America, Middle East/ North Africa. Projects that operated in multiple regions are assigned 'Cross-regional'.	Evaluation reports.
Time between final review and EPE	Discrete variable. Time between final review and ex post evaluation (financial sector: year of last disbursement). In case the final review is dated after the evaluation year (33 instances), the value is set to zero. Unit: years	Evaluation reports.
Year of project start	Categorical variable. 5-year intervals of year of project start. First interval: 1990–1994. Last interval: 2015–2019.	KfW data.
Year of evaluation	Categorical variable. 5-year intervals of year of ex post evaluation. First interval: 2005–2009. Last interval: 2020–2024.	Evaluation reports.
Analytical variables		
Report weight	Inverse of number of projects evaluated in the same report.	Evaluation reports.
Exclude	Observations excluded from analyses: 1. Project start prior to year 1990 (n=3) 2. Interim evaluations (n=1) 3. "Promotional Loans" (n=4) 4. Evaluations commissioned by Federal Foreign Office (AA) (n=5) 5. Canceled projects (n=6) 6. Projects operating in 'all developing countries' (n=4) or across a large number of countries (n=1)	-

A.2 Figures

Figure A1: Distribution of DAC-ratings



Note: Distribution of the overall and individual DAC-ratings for $N = 1,124$ evaluations in our sample. Frequency refers to the number of projects with respective rating. Red line depicts the normal density.

A.3 Methodology

A.3.1 Calculation of macro variables

All macro variables we employ in our model are computed as averages using the respective variable's average value a) during its project duration and b) across countries in which the project was implemented.

$$macro_i = \frac{1}{N} \sum_{k=1}^{N} [\frac{1}{T-t} \sum_{s=t}^{T} macro_{kt}] \tag{2}$$

where i denotes the respective project, N is the number of countries in which the project was implemented in, T is the year in which a project was completed and t is the year a project started.

A.3.2 Extra-/Interpolation for macro variables

The macro variables used are partially not available for the relevant time period. We impute these missing values in two steps:

Annual growth rate:

$$agr_t = var_t / var_{t-1}, \tag{3}$$

where var_t is the variable of interest in year t and agr is the annual growth rate.

Geometric mean:

$$geomgr_t = (agr_t \cdot agr_{t-1} \cdot agr_{t-2} \cdot agr_{t-3} \cdot agr_{t-4})^{1/5}, \tag{4}$$

where $geomgr_t$ is the geometric mean of the 5 last annual growth rates (last 5 years) in year t.

i) If latest data points are missing (forward extrapolation) or if data is missing in between available data points (interpolation):

   $variable_t = variable_{t^*} \cdot geomgr_{t^*}$

   t^* corresponds to the latest year for which the variable is available (and its five-year geometric growth rate). $t^* < t$.

ii) If earliest data points are missing (backward extrapolation):

   $variable_t = variable_{t^*} \cdot geomgr_{t^*}$

   t^* corresponds to the earliest year for which the variable is available (and its five-year geometric growth rate). $t^* > t$.

In case macro data observations are missing for certain years, we use the five-year geometric mean of annual growth rates to extrapolate (interpolate) missing values.

A.3.3 Within- vs. between-country analysis

We regress binary project success on a categorical indicator for the the country a project was implemented in to determine whether success varies more within-country or between-country.[12] This regression is run separately in sub-samples for each year projects in our sample were active in, i.e. 1990-2020.[13] We define active in year t as project start prior to or in t and project end before or in t. The following equation describes the model:

$$y_{ti} = \alpha_t + \beta_t \, country_i, \tag{5}$$

where y_{ti} is a binary measure of success of project i in sub-sample t and $country_i$ is a categorical variable, corresponding to the country project i was implemented in.

To obtain a single numeric value across sub-samples, we take the weighted average of regression coefficients. The weight corresponds to the number of observations in each sub-sample:

$$\hat{\beta} = \frac{\sum_{t=1}^{T} \hat{\beta}_t \cdot N_t}{\sum_{t=1}^{T} N_t}, \tag{6}$$

T is the number of years a project was active in and $\hat{\beta}_t$ is the parameter for projects active in year t and N_t is the number of observations (projects) active in year t.

The explanatory power of *between*-country variation in project success is evaluated as follows: The estimated parameter $\hat{\beta}$ is interpreted as the explanatory power of between-country variation. The unexplained variation, $1 - \hat{\beta}$, is interpreted as within-country variation, i.e. project-specific characteristics.

Note that some projects ($N = 87$) were implemented in multiple countries. For these countries, a country group was created and entered as country. A possible caveat of this procedure is that the explanatory power of between-country variation may be inflated: If in a specific set of countries, only one project of our sample was active, the country or country group explains 100% of the variation in outcomes (cf. Bulman et al. (2017)). In our dataset, this only applies to two observations.

[12] Or countries, depending on the number of countries a project was implemented in.
[13] Because only one project was active was active in 2021, we exclude this year.