Courant Research Centre
'Poverty, Equity and Growth in Developing and Transition Countries: Statistical Methods and Empirical Analysis'

Georg-August-Universität Göttingen
(founded in 1737)

Discussion Papers

No. 154

Analyzing Effectiveness of Development Aid Projects: Evaluation Ratings or Project Indicators?

Laura Metzger, Isabel Günther

November 2013

Wilhelm-Weber-Str. 2 · 37073 Goettingen · Germany
Phone: +49-(0)551-3914066 · Fax: +49-(0)551-3914059

Email: crc-peg@uni-goettingen.de Web: http://www.uni-goettingen.de/crc-peg

Analyzing Effectiveness of Development Aid Projects:
Evaluation Ratings or Project Indicators?

Laura Metzger
ETH Zürich

Isabel Günther[1]
ETH Zürich

Abstract

Ongoing empirical research on the drivers of project-aid effectiveness relies on World Bank *evaluation ratings* across heterogeneous aid sectors. This leads to two problems. First, it is difficult to identify which dimension of project performance World Bank evaluation ratings are measuring precisely. Second, only *project management* variables, which are sector independent, can be included in the analysis. This study concentrating on an analysis of 150 water supply projects enables us to work with a more precise and objective performance measure by defining sector-specific *indicators of improved water supply*. Moreover, we are able to analyze the impact of *project design* in addition to project management. We find that evaluation ratings and indicators of improved water supply are positively but weakly correlated. Project management variables have a higher impact on evaluation ratings whereas project design variables have a higher impact on improving water supply to the target group. Various independent variables even change sign if indicators of improved water supply instead of evaluation ratings are chosen as a performance measure of project-aid effectiveness. Taking into account project design in addition to project management and country characteristics considerably increases the share of variation in project performance that can be explained.

JEL Code: F35; O19.
Key Words: Aid effectiveness; evaluation; indicators; water supply.

Acknowledgement: We thank the German Development Bank (KfW) for supporting our research and making their project and evaluation documents available to us. We especially thank Eva Terberger and Jan Schrader for highly valuable comments on earlier drafts of this paper and very interesting discussions. Any errors remain the sole responsibility of the authors.

[1] *Corresponding Author*: Isabel Guenther, ETH Zurich, Center for Development and Cooperation, Clausiusstrasse 37, 8092 Zurich, Switzerland. Email: Isabel.guenther@nadel.ethz.ch. Phone: +41 44 6328750. Fax : +41 44 6321207.

1. Introduction

The effectiveness of development aid is one of the most debated and most important issues in development research and policy. The majority of research concentrates on studying aid effectiveness either at the macro (see e.g. Burnside and Dollar, 2000 and 2004; Collier and Dehn, 2001; Collier and Dollar, 2002; Dalgaard, Hansen and Tarp, 2004; Clemens et al., 2004; Roodman 2007) or the micro level.[2] Research on aid effectiveness at the project level is scant, because project related data are not readily accessible to the public. Existing studies on the effectiveness of project-based aid are almost exclusively confined to World Bank projects. The World Bank dataset is the only publicly available database on projects across countries including a measure of project success. Project success is indicated by the World Bank's or the Independent Evaluation Group's (IEG) project evaluation ratings (e.g. Denizer et al., 2013; Dollar and Levine, 2005; Guillaumont and Laajaj, 2006; Dreher et al., 2013).

Using these evaluation ratings to analyze (the correlates of) aid effectiveness is, however, problematic because of the following: First, the ratings are the result of the specific evaluation system (and rules) that a donor applies for the assessment of its projects. Donors may differ in their views as to what constitutes a successful project and therefore set up their evaluation criteria accordingly. Second, evaluation ratings are a product of an evaluator's subjective assessment of project performance. Although there are rules and indicators that evaluators have to consider when evaluating a project, they are given a fair margin of discretion in their rating decisions so that they are flexible enough to take account of a project's individual context. Moreover, the overall evaluation ratings are rather complex indicators as they are often composed of several "sub-ratings" that capture different dimensions of a project which are considered important for its success. It is therefore difficult to precisely identify what the ratings are measuring as they combine different criteria into a single judgment.

Hence, evaluation ratings of projects are not directly comparable between, and not even necessarily within, donor organizations. Most studies intensify these problems as they aggregate the ratings over heterogeneous aid sectors, thereby further diluting their comparability and informative value. An additional drawback of the aggregation approach is that it limits the range of project specific variables from which researchers can choose to explain project outcomes, as they are restricted to sector-independent variables. As a consequence, studies analyzing aid effectiveness across projects apply a vague definition of project performance, use a limited set of *project management* variables that are available for all sectors (e.g. Denizer et al., 2013; Kilby, 2000), and hence explain only a small share of the variation in project performance.

[2] The examples are numerous given that almost every aid intervention has to be evaluated separately. For an overview of randomized controlled trials, having become the most widely applied method to study aid effectiveness at the micro level, see for example www.povertyactionlab.org. For an overview of studies summarizing micro-level aid effectiveness studies by sector see for example www.3ieimpact.org.

This study of 150 drinking water supply projects financed by the German Development Bank (KfW), on behalf of the German Government, seeks to contribute to the literature on project aid effectiveness in the following way. Instead of pooling evaluation ratings of different sectors, we exclusively concentrate on (all) drinking water projects (financed and evaluated by KfW over 50 years). This enables us to work with a more precise and objective measure of project performance than previous studies. Project design (and not management) is usually the focus of research on aid effectiveness at the micro level, with randomized controlled trials (RCTs) currently being the most common method. We define verifiable indicators that are comparable across all water supply projects. Moreover, we are able to consider a broader range of project level variables: in addition to country characteristics and general *project management* variables we are able to include and analyze sector specific *project design* variables. This allows us to bridge the gap between macro (country characteristics), project (project management), and micro (project design) studies on aid effectiveness.

We chose water aid, because it is a sector with quantifiable goals and indicators that are uniformly applied to all of KfW's water projects (KfW, 2009a and 2009b) and that could, in principle, also be applied to other donors' drinking water projects. A further advantage of the water sector is that project goals have changed very little over time. Hence, the chosen indicators are valid for all years and all projects in the sample. Moreover, drinking water aid was, and still is, one of KfW´s main fields of activity. On behalf of the German Government, the organization currently finances drinking water and sanitation projects in more than 50 countries, targeting more than 41.3 million people.[3] In 2011 German Development Cooperation (including KfW and other organizations) was, together with Japan, the largest bi-lateral donor in the water and sanitation sector (Anand, 2013). Germany alone accounted for 21.5 percent of total aid worldwide given to the water and sanitation sector.

The water access indicators we use to measure project performance – in addition to, and in comparison to, evaluation ratings – are: (i) *safe water consumption (in liters) per capita and day* and (ii) *the number of individuals with access to an improved water source*. The two indicators are stated in the official results chain[4] of KfW's drinking water projects (KfW, 2009a and 2009b). The respective values were extracted from the project documents of the 150 projects we analyzed. They measure if, and to what extent, the immediate project goal - which is to provide the target population with a sufficient amount of safe drinking water - was achieved at the time of evaluation.

We are aware that these indicators are narrowly defined and not directly comparable to the evaluation ratings that go beyond mere supply targets to reflect project success. We nevertheless

[3]https://www.kfw-entwicklungsbank.de/Internationale-Finanzierung/KfW-Entwicklungsbank/Sektoren/Wasser/Engagement-der-KfW-Entwicklungsbank/

[4] The results chain or "logical framework" stipulates how a project's inputs and activities lead to the desired key objectives. Each key objective is linked to a defined set of indicators.

deem them adequate alternatives to measure project performance. First of all, they reflect the project's immediate benefit for the target group in terms of both the quantity of water delivered and the number of beneficiaries reached. Moreover, an improved drinking water supply is a necessary precondition for other positive, indirect, welfare effects the projects are targeting, such as reduced time and physical burden for water fetching, and improved health as a result of reduced water-borne diseases. Secondly, these indicators should not be at odds with the evaluation ratings, i.e., the relationship between ratings and objective indicators should be positive. Third, and in contrast to evaluation ratings, these indicators are not donor-specific, but are in line with internationally agreed drinking water indicators as stipulated in the Millennium Development Goals[5] (MDGs) and the United Nations Declaration of the Human Right to Water.[6] More specifically, the indicator of target 7C[7] of the MDGs measures the "proportion of population using an improved drinking water source".[8] The UN's *Right to Water* declaration demands that a nation's entire population is provided with access to safe water and that "the water supply for each person must be sufficient and continuous to cover [*basic*] personal and domestic uses", which is said to be guaranteed with a consumption of 50 to 100 liters per person and day.[9]

We find that evaluation ratings and the objective indicators of improved water supply are positively but weakly correlated. Moreover, our findings suggest that the explanatory variables affect project performance differently, depending on whether it is measured with the evaluation ratings or the project indicators, and that this leads to different conclusions about the drivers of increased aid effectiveness. General project management variables, such as *project duration* or *perceived risk of project failure*, have a higher impact on evaluation ratings. Specific project design variables, such as the type of water supply system or the existence of established water user committees, have a higher impact on indicators of improved water supply. Similar to other studies we find that general country level characteristics explain little of the variation in project outcomes. Taking into account project-level variables increases the explanatory power of our models considerably.

The rest of the paper is organized as follows. In the next section we present the literature on effectiveness of development aid projects. In section 3 we describe our data and the methods used. In section 4 we present and discuss the results; section 5 concludes.

[5] http://www.un.org/millenniumgoals/environ.shtml.
[6] See *Right to Water: Human Rights Fact Sheet No.35*, United Nations (2010).
[7] Target 7C reads: "Halve, by 2015, the proportion of the population without sustainable access to safe drinking water and basic sanitation."
[8] http://mdgs.un.org/unsd/mdg/host.aspx?Content=indicators/officiallist.htm.
[9] See p.8 of *Right to Water: Human Rights Fact Sheet No.35*, United Nations (2010).

2. Literature Review

Studies on project-level aid, most of them cross-sectional, can be roughly divided into two strands of literature. Studies in the first strand strictly focus on country characteristics to explain World Bank project performance (see e.g. Dollar and Levine, 2005; Guillaumont and Laajaj, 2006; Isham, et al., 1997; Isham and Kaufmann, 1999; Dreher et al., 2013). Studies in the second strand add, to a varying degree, project management variables to the set of country characteristics to explain project performance (see e.g. Denizer et al., 2013; Dollar and Svensson, 2000; Kilby, 2000). Most of the above studies either use World Bank and/or IEG (formerly OED, Operations Evaluation Department) [10] project evaluation ratings as an indicator for project success. Only Hemmer and Lorenz (2003) focus on KfW projects and use KfW's project ratings to measure project performance (see section 3), and Ivanova et al. (2001) analyze International Monetary Fund (IMF) programs using program interruptions, compliance with conditionality, and the share of committed funds disbursed as project performance indicators. All studies aggregate the project ratings over different aid sectors. We are not aware of any study that takes a sector specific approach to analyze the effectiveness of development aid projects.

The IEG and the World Bank use a common project evaluation approach. World Bank project managers as well as IEG evaluators assess if (i) a project achieved its stated objectives (outcomes), (ii) if the outcomes are sustainable, (iii) if resources were used efficiently in achieving the objectives, and (iv) if the project's design and objectives were adequately tailored to the development problem in question. Since 1995, a six-point project rating scale from "highly unsatisfactory" to "highly satisfactory" has been used. Before 1995, the rating scale was binary, which allowed projects to be categorized into "satisfactory" or "unsatisfactory." The IEG evaluations are conducted by individuals that were never involved in the project being evaluated, while evaluations at the World Bank are done by individuals who have implemented and/or managed the project under consideration. For further details see for example Guillaumont and Laajaj (2006), Denizer et al. (2013), Dreher et al. (2013) as well as the IEG's website[11] and the website of the World Bank.[12]

Studies focusing on country-level determinants of project performance can be briefly summarized as follows. According to Dollar and Levine (2005), high-quality institutions significantly raise the probability for a project to be successful, even in low-income countries. When disaggregating the success ratings by investment sector, they find that the impact of institutional quality varies highly for different sectors. This result highlights that it might be

[10] The OED was founded in 1970 as the World Bank's first independent evaluation unit. In 2006 it was merged into the newly founded IEG.
[11] http://ieg.worldbankgroup.org/ratings
[12] http://web.worldbank.org/WBSITE/EXTERNAL/PROJECTS/EXTPOLICIES/EXTOPMANUAL/0,,contentMDK:20064672~menuPK:64701637~pagePK:64709096~piPK:64709108~theSitePK:502184~isCURL:Y,00.html

important to differentiate between (heterogeneous) aid sectors. In line with these results, Isham et al. (1997) find that higher levels of civil liberties (e.g. freedom of individual expression, a pluralistic and free media, the ability of groups to organize, and freedom of dissent and criticism) are significantly correlated with better project performance. A complementary study by Dreher et al. (2013) examines if "politically motivated aid", i.e. aid that is allocated for political reasons and not because of need, is less effective. They find that project performance is only negatively affected by politically motivated aid if the recipient country is economically vulnerable. Otherwise politically motivated aid does is not evaluated worse than aid that is not politically motivated.

According to Guillaumont and Laajaj (2006), economic instability - approximated by the variability of a country's exports - reduces the probability for projects to be successful. Moreover, they find that the probability of success also decreases with an increasing amount of aid received. The latter result parallels the diminishing-returns-to-aid finding from the macroeconomic literature (Dalgaard, Hansen and Tarp, 2004) which is typically explained by the limited absorptive capacity of recipient country institutions. Last, Hemmer and Lorenz (2003) find that projects tend to be less successful in Middle East and Northern Africa and more successful in Asia and Latin America and the Caribbean. Furthermore, they find that a better macroeconomic environment (less inflation, balanced budget) and trade openness are significantly correlated with better project performance.

In addition to country level characteristics, Dollar and Svensson (2000) analyze a set of project management variables, such as the number of conditions attached to a loan as well as project preparation and supervision time. They find that their set of project variables has no explanatory power, while politics - democracy and government crisis during reform periods - are significantly related to project performance. They conclude that the World Bank has little influence on project outcomes via monitoring and supervision in a politically and economically risky environment. Similarly, Ivanova et al. (2001) find that for structural adjustment programs of the IMF, political instability, political cohesion, and influence of political interest groups in recipient countries matter for project success, while the IMF's project steering and management efforts have no significant effect on project outcomes. Also Kilby (2000) investigates how supervision time affects the success of World Bank projects. In contrast to Dollar and Svensson (2000) and Ivanova et al. (2001), Kilby (2000) finds that additional supervision time, especially when invested early in the project management process, is significantly correlated with a higher project rating. While Kilby uses World Bank evaluation ratings to measure project performance, Dollar and Svensson (2000) use the OED ratings.

A study by Denizer et al. (2013) is, to our knowledge, the most comprehensive study in this field so far. For the project-related variables they find that early warning indicators, which are issued by task managers to mark potentially risky projects, are significantly negatively correlated with project performance ratings. Furthermore, their data show that larger projects and projects with a longer lag between project closure and evaluation tend to get lower ratings. Last, the quality of the task manager is significantly positively correlated with project ratings. With respect to country characteristics, they find that evaluation ratings are strongly positively correlated with country CPIA (Country Policy and Institutional Assessment) scores. The CPIA index is published by the World Bank and indicates if a country's institutional and policy framework is "[…] conducive […] to fostering poverty reduction, sustainable growth, and the effective use of development assistance".[13] This result supports the findings of Dollar and Levine (2005). In 2008, KfW's independent evaluation department conducted a short (and unpublished) analysis on the relationship between country- and project-level characteristics and the performance of KfW projects (KfW, 2008). The study shows that a longer lag between project closure and evaluation reduces evaluation ratings and that projects which KfW task managers initially marked as risky have a lower success probability. Both results parallel the findings of Denizer et al. (2013).

Common to all studies is the fact that country characteristics and the project management variables only explain a small (to moderate) share of the variation in project performance (evaluation ratings). The pseudo R2s of Guillaumont and Laajaj's (2006) ordered logit regressions are below 0.1. Isham et al. (1997) attain R2s between 0.02 and 0.18, while the pseudo R2s of the regressions in Dreher et al. (2013) vary between 0 and 0.09. The explanatory power of the main probit estimations in Hemmer and Lorenz (2003) shows pseudo R2s between 0.06 and 0.15. The R2s of the linear probability models of Dollar and Svensson (2000) range between 0.15 and 0.28. Last, Denizer et al. (2013) find that country characteristics account for around 10 percent of the total variation in evaluation ratings, while project-related variables account for about 5 percent.

3. Data and Empirical Approach

3.1 Background of Data Set

KfW´s independent evaluation department (FZ E hereafter) is responsible for conducting and supervising project evaluations. Before 2007, the evaluation department evaluated every single completed project. Since 2007, the internal evaluation policy has stipulated that 50 percent of all completed projects be evaluated 3 to 5 years after their finalization. KfW uses a stratified random sampling approach to generate a representative sample of the projects which are to be evaluated. Evaluations are usually based on a review of the relevant project documentation and a two-week field trip to the project site, including personal interviews with beneficiaries, the local project partners, other partner organizations, and higher ministries involved in project steering and implementation (KfW, 2011b). Roughly 10 to 15 percent of yearly evaluations are commissioned to external consultants. The rest are carried out by KfW staffs who were not

[13] http://www.worldbank.org/ida/papers/CPIAcriteria2011final.pdf

involved in the management of the project they are assigned to evaluate, in order to avoid a conflict of interest. As far as drinking water projects are concerned, a civil engineer sometimes joins the evaluation team to provide technical expertise in case it is needed to assess the functionality of the installed water supply system. The final evaluation report is a qualitative assessment of project performance, supplemented with descriptive, quantitative data.

Our dataset comprises 150 KfW financed drinking water projects, distributed over 62 countries and spans the years between 1965 and 2011. The list of countries can be found in Table A1 in the Appendix. We consider all drinking water projects that FZ E has evaluated between 1989 and 2011, and for which (hard-copy) documentation is available.[14] Ex ante and ex post values for the *project management and design* (see Section 3.4) variables and for the *project performance indicators* (see Section 3.3) were extracted by hand from hard-copy documentation. This required a detailed reading of a total of 150 project appraisals and 150 evaluation reports. A single report is, on average, about 50 pages. As explained in more detail below, the project appraisal reports provided us with baseline values. Project *evaluation ratings* were taken from FZ E's database (we cross-checked its entries with the entries in the hard-copy project reports).

3.2 Performance Measure 1 - Evaluation Ratings

In 2006, FZ E aligned its evaluation criteria with the OECD-DAC guidelines and since then has evaluated all projects with respect to five criteria*: relevance*, *efficiency*, *effectiveness*, *impact* and *sustainability*. Between 2000 and 2006, it evaluated all projects on the basis of their *significance* (similar to relevance), *efficiency*, *effectiveness* and *impact*. Although *sustainability* did not exist before 2006 as a stand-alone criterion, it was treated as a cross-cutting issue in project evaluations. Before 2000, evaluation reports only recorded an *overall* project rating.

The direct project goal, or outcome, of drinking water projects is to supply a defined target population with a sufficient amount of safe drinking water (*effectiveness*).[15] This is achieved by means of a real investment in water supply infrastructure. The direct project outcome is considered a necessary condition for achieving the overall project goal, which is to improve the target population's health by reducing and preventing waterborne diseases (*impact*), ideally in the long-term (*sustainability*). In order to achieve these goals, it is considered vital that the water supplier is able to cover the full costs[16] of drinking water provision, preferably out of tariff revenues (*efficiency*).[17] It is argued that full cost coverage is necessary to guarantee the target

[14] Note that our dataset starts in 1965 although project evaluations started in 1989 only. Some projects that started in the 1960s had such a long lifespan that they were evaluated in or after 1989. The oldest project in the sample started in 1965 and was evaluated in 1991.

[15] Safe means the water quality should comply with WHO guidelines: http://www.who.int/water_sanitation_health/publications/2011/dwq_chapters/en/index.html

[16] Full costs comprise the costs for system operation and maintenance as well as (re)-investments cost in the physical infrastructure. The precise definition according to KfW (2008a) reads: "The full costs are defined as…operating costs (staffing, operation, maintenance, energy, etc.) + depreciation + appropriate return on the capital employed."

[17] Cost coverage is only one component entering the efficiency criterion. For further details see:

group´s sustainable supply with sufficient drinking water in the long-term.[18] In the short-term, being able to cover operating costs is seen as acceptable (KfW, 2008a). If the water supplier is servicing debt, "revenue should cover at least the operating costs and debt servicing" (ibid). *Relevance* is used to validate first, if the project in question targeted an important development problem in the partner country at the time it was designed and, second, if its development goals are in line with those of the partner country, the German Government, and international development standards (e.g. the MDG's and the Paris Declaration).

The evaluation rating scheme itself works as follows. Every criterion is first rated separately and then combined into a single *overall project performance rating*. *Relevance*, *impact*, *effectiveness* and *efficiency* are rated on a 1 ("very successful") to 6 ("project failed") scale. By contrast, *sustainability* is rated on a 1 ("very good sustainability") to 4 ("insufficient sustainability") scale, because, according to FZ E, there is no increase to "insufficient" sustainability. The five separate ratings are then aggregated for a final rating which reflects the project´s overall performance. Similar to the World Bank's evaluation rating, no rule prescribes the weighting of the separate criteria when combining them into an overall rating. However, if a project "fails" with respect to any of the criteria *effectiveness*, *impact* or *sustainability*, the entire project is categorized as unsuccessful (KfW, 2006b and KfW, 2010). The scale of the final rating also ranges from 1 ("very successful") to 6 ("project failed"). Projects rated between 1 and 3 are considered successful; those rated higher (i.e. worse) than 3 are considered unsuccessful. In order to facilitate the interpretation of coefficients in Section 4, we reverse the rating scale, with 1 referring to "project failed" and 6 referring to "very successful". Thus projects rated 4 and higher are considered successful, whereas those rated 3 and lower are considered unsuccessful.

Following previous studies (see Section 2) we first of all measure project performance by means of the *overall* evaluation ratings. In addition, we also analyze the *effectiveness* ratings. We are particularly interested in a project's *effectiveness*, because this criterion is supposed to reflect whether the immediate project goal - to supply a defined target population with a sufficient amount of safe drinking water - was achieved or not. Consequently, the *effectiveness* rating is explicitly linked to the project indicators described in next section, i.e. *water consumption per capita in liters* and *the number of project beneficiaries with access to an improved water source*. Of the 150 projects considered here, more than 98 per cent of the projects explicitly aimed to provide water to more individuals and 92 percent aimed to provide those individuals with a larger quantity of drinking water. Data on drinking water quality are usually not available in the projects' documentation. Hence we were not able to consider *water quality*. Sufficient data on

https://www.kfw-entwicklungsbank.de/Internationale-Finanzierung/KfW-Entwicklungsbank/Evaluierung/Ex-Post-Evaluierungen/Schlüsselkriterien/

[18] According to the guidelines set by the German Ministry for Economic Cooperation and Development, full cost coverage via tariffs is the way to ensure a sustainable (i.e. long term) drinking water provision. State-subsidized full cost coverage is acceptable, but should be the exception to the rule. See also: http://www.bmz.de/de/publikationen/reihen/strategiepapiere/Konzepte143.pdf.

the target group´s health situation before and after an intervention are also not available, so that we neither consider a project's *impact* and *sustainability* in our analysis.

3.3 Performance Measure 2 - Project Performance Indicators

For each project, appraisal reports are prepared shortly before a project actually starts. These reports provide us with baseline values, i.e. with actual supply levels (t0_baseline), and with planning values for (t1_target) on daily *water consumption per capita in liters* and on *the number of project beneficiaries with access to an improved water source.* Together with the actual values of these indicators recorded at the time of evaluation (t1_actual), we are able to construct a series of three variables that allow us to account for the possibility that, when evaluating a project, evaluators make a conceptual difference between the following three dimensions of project performance: (1) total supply levels at evaluation, i.e., the total number of individuals with access to an improved water source and the absolute quantity of safe water they consume; (2) the change in supply levels between appraisal (the baseline) and evaluation, i.e., the change in the number of individuals with access to an improved water source and the change in the quantity of safe water they consume between t0 and t1; and (3) the goal achievement at t1, i.e., the extent to which the actual supply levels (t1_actual) met the targeted supply levels (t1_target) at the time of evaluation (see Table 1).

Table 1 Definitions of project performance indicators

Objective performance indicators	Definition
(1) Total supply levels at t1	
water pcl	log of daily safe water consumption per capita in liters (pcl) at t1
population	log of number of people in the target population with access to an improved water source in t1
(2) Change in supply levels t1-t0	
change water pcl	relative (percentage) change in daily safe water consumption per capita in liters between t0 and t1
change population	Relative (percentage) change in number of people in the target population with access to an improved water source (number of people provided with an improved water source by the project) between t0 and t1
(3) Goal achievement at t1	
goal achievement water pcl	actual divided by targeted daily safe water consumption per capita in liters in t1
goal achievement population	actual divided by targeted number of people with access to an improved water source in t1

We expect a conceptual difference between the above three versions of the objective indicators, because judging the success of a project either on the grounds of total supply levels at the time of evaluation, or a before-after change in supply levels, or the extent to which targets were achieved are different things. In fact, total supply levels take stock of the target group's situation at evaluation (t1), but put little emphasis on the impact of the project. By contrast, before-after changes focus on the benefits the project generated for the target group and should be important for evaluators. Looking at a project's target fulfillment is a way of judging its success with a stronger focus on project planning.

3.4 Explanatory Variables

Country-level characteristics are taken into account with the following set of variables: GDP per capita is used as an indicator for the country's development status and inflation indicates its macroeconomic stability. Both indicators are taken from the World Development Indicators (2013). The absolute and squared value of bilateral aid per capita, also taken from the World Development Indicators (2013), are supposed to capture whether the total level of bilateral aid in the country has an influence on project outcomes and whether there is, as it is often claimed in the literature, a diminishing return to aid. We computed the yearly average of GDP per capita, bilateral aid per capita, and inflation rate over the length of the project, i.e. from its start to its evaluation.

We further create a dummy variable from the UCDP/PRIO data on armed conflict, which is 1 if there was a minor conflict (25-999 battle related deaths) and 2 if there was a war (over 999 battle related deaths) at any point in time between project start and evaluation and 0 otherwise.[19] The potential influence of the regime type on project success is captured by the *polity2* indicator by Marshal et al. (2013)[20] which ranges from -9 to 10. Higher scores represent a more democratic regime. We use the *polity2* indicator at project start in our regressions. We also consider a region fixed effects for sub-Saharan Africa (being the poorest region) and a region fixed effect for the Middle East (being the region with the highest water scarcity). Region effects are preferred to country fixed effects to save degrees of freedom. Our set of variables is similar to the one used in the study by Denizer et al. (2013). In addition to their variables, we consider bilateral aid and conflict. We refrain from further extending the set of country characteristics, as this would result in a reduction in sample size.

As explained above, project-level variables were extracted from hard-copy project documentation records. First, we compiled a list of, what we call, *project management* variables that are similar (and comparable) across aid sectors. These variables are similar to the type of variables used in previous studies (e.g Denizer et al., 2013). *Project size* is reflected by total project costs. We also account for the *recipient share in total project costs*. This variable can be interpreted as an indicator for the general financial capacity of the partner country, but may also indicate the importance drinking water issues have in its policy agenda. In case this holds true, we would observe a positive relationship between the recipient's share in total project costs and the performance indicators. We also consider the *project duration* as well as the *time elapsed between project closure and evaluation* (measured in months). Previous studies found that higher values of these variables are negatively correlated with project performance (Denizer et al.,

[19] http://www.pcr.uu.se/research/ucdp/datasets/ucdp_prio_armed_conflict_dataset/
[20] http://www.systemicpeace.org/polity/polity4.htm

2013). Furthermore, we consider the *phase of a project* ranging from the first phase to the third phase. Project phases correspond to separate projects with separate financing and a separate set of project goals. Finally, the project manager's *perceived risk of project failure* is captured by a variable that takes values between 1 ("low risk") and 4 ("very high risk"). Previous studies based on KfW and World Bank data found the risk indicators to be negatively correlated with evaluation ratings.

In addition to *project management* variables, we consider a number of *project design* variables that are specific to the water sector. Microeconomic studies on aid effectiveness usually focus on understanding what type of project design improves aid effectiveness, whereas macro-economic studies focus on country characteristics (of recipients and donors), and cross-sectional studies on project-level data has focused on project management (or country characteristics). At least to our knowledge, we are the first to combine all dimensions in one study.

We study whether the project was located in *an urban area*; whether a *water user committee* was put in place to operate and maintain the supply system; whether there was an activity to *institutionally strengthen the water supplier* and/or some other body responsible for improving the drinking water supply in the project region; whether a *wastewater* component was integrated into the program; whether a *hygiene campaign* was part of the program's activities; and which type of *supply system* is used to supply the target population with safe drinking water: non-piped, mixed (piped and non-piped), or piped system. Another important sector-specific variable we consider is the local tariff that is charged per meter cubed of consumed water. Last, we consider a dummy variable that indicates whether a country has low internal water resources. This last variable is generated from data provided in the FAO aquastat databank.[21]

An overview of all explanatory variables and project performance variables is provided in Table A2 in the Appendix.

3.5 Estimation Strategy

Our estimation strategy is divided into two parts. First, we analyze the correlation between the *overall* rating and its five constituent elements: *effectiveness*, *efficiency*, *impact*, *sustainability* and *relevance* (see Table 2). Next, we examine the correlation between the *overall* and *effectiveness rating* and the project indicators (see Table 3). *Effectiveness* is used as an additional rating next to the *overall* rating, since it is a key evaluation criterion and because it is directly linked to the project indicators. Moreover, we expect conceptual differences between the *overall* and the *effectiveness* ratings, because the former unifies all dimensions of a project that are considered important for its success, while the latter is more narrowly defined as it concentrates on measuring the immediate project outcome. The aim of this first part of the analysis is to analyze the correlation between the *overall* project rating and its sub-ratings (*relevance*,

[21] http://www.fao.org/nr/water/aquastat/data/query/index.html

effectiveness, efficiency, impact, and *sustainability*) on the one hand, and the correlation between evaluation ratings and the objective project indicators on the other hand.

In the second part of our analysis, we seek to test if there is a conceptual difference between the project indicators and evaluation ratings. We study (a) how much of the variation in project performance is explained by country-level characteristics and project-level characteristics and (b) how the results vary with the indicator that is used to measure project performance (see Table 4-6).

We work with a pooled cross-section of 150 projects across time. Given that the sample of annually evaluated projects comprises either the entire set of finalized projects or, since 2007, a randomly drawn sample thereof, we assume to be working with an independently pooled cross-section. Due to missing data, especially for the baseline values of number of people with access to an improved water source and water quantities consumed, the sample size varies (and is mostly smaller than the initial sample of 150 projects) based on the specification chosen. The data is, however, missing at random (test statistics are available from the authors on request). Nevertheless, the small sample size may decrease the precision of estimates as confidence intervals become wider and standard errors larger, and regression results can be sensitive to changes in explanatory variables. Yet, it should be noted that similar studies operate with comparable or even smaller sample sizes: Dollar and Levine (2005) for example work with a sample size of 52 to 90 observations; Ivanova et al. (2001) work with a sample size of 55 to 61. Moreover, most results are robust to the specification chosen (see Table 4-6 and Table A4 in the Appendix).

We are aware of the fact that our results rely on observational data and are based on a cross-sectional data analysis. Causal inference is hence only possible to a limited extend. However, previous literature in the field has faced the same problems (and we compare our results to this literature). Moreover, we focus on comparisons and differences between various specifications (rather than identifying drivers of aid effectiveness). Hence, problems of endogeneity should be less severe.

We estimated all regressions with ordinary least squares (OLS) regressions. In addition, all specifications with evaluation ratings were also estimated with ordered logit models, given that evaluation ratings are rank-ordered categories. There is a long standing debate among econometricians about whether it is justified to use linear regression models for categorical outcomes, or whether this approach produces biased estimates (Long, 1997). Despite this qualification, only OLS estimates are presented here.[22] First, OLS is often applied to categorical outcome data, because it greatly simplifies the interpretation of regression coefficients. Moreover, results from evaluation ratings are easier to compare to specifications with objective

[22] Ordered logit estimations are available from the authors on request.

project indicators (which are continuous outcome variables). Second, as far as our empirical analysis is concerned, the results do not change in terms of sign and significance when we estimate the regressions with an ordered logit instead of an OLS model. Last, also Denizer et al. (2013) use OLS for binary and categorical outcomes throughout their entire analysis in order to simplify the interpretation of the point estimates in their regressions.

Last, even though we discuss some details of drinking water projects when discussing our results in next section, the focus of this analysis is on the problem of measuring project performance, not on water aid. An in-depth discussion of water aid is beyond the scope of this study.

4. Results

Columns (1) to (5) of Table 2 show that each of the partial evaluation ratings (*effectiveness, efficiency, impact, relevance and sustainability)* is, as expected, significantly positively correlated with the *overall* evaluation rating. Hence, the better a project performed in any of the OECD-DAC dimensions, captured by the partial ratings, the higher the *overall* rating. The regression presented in column (6) of Table 2 correlates the *overall* performance ratings simultaneously with all partial ratings. For this analysis, we have to exclude *sustainability* because the number of observations would be reduced to 32 otherwise. We observe that *relevance* turns insignificant once analyzed jointly with the other ratings. This result implies that *relevance* does not exert a strong influence on overall project evaluations. We argue that this result is not surprising because projects that are not relevant to recipient countries' development and/or to German Development Cooperation would not be undertaken. Furthermore, we observe that *effectiveness* has the strongest marginal impact on the *overall* rating. This is in line with the results from the partial regressions, where effectiveness alone explains almost 80 percent of the variation in the overall rating, whereas other dimensions only explain between 50 and 60 percent.

Table 2 Correlation between partial ratings and overall evaluation rating

	Overall Evaluation Rating					
	(1)	(2)	(3)	(4)	(5)	(6)
effectiveness rating	0.875***					0.454***
	(0.000)					(0.000)
efficiency rating		0.736***				0.238***
		(0.000)				(0.000)
impact rating			0.827***			0.384***
			(0.000)			(0.000)
sustainability rating				0.812***		
				(0.000)		
relevance rating					0.827***	-0.041
					(0.000)	(0.620)
Constant	0.431**	1.230***	0.480*	0.820	0.267	-0.214
	(0.036)	(0.000)	(0.073)	(0.170)	(0.438)	(0.267)
R2	0.785	0.605	0.686	0.488	0.588	0.887
Adjusted R2	0.782	0.600	0.683	0.473	0.584	0.881
N	93	93	89	35	91	86

Notes: Standard errors in parentheses, *significant at 10%, **significant at 5%; *** significant at 1%; see Table 1 for precise definitions of dependent variables.

Because KfW's evaluation criteria were aligned with the OECD/DAC criteria in 2006, it would be interesting to check – by considering *sustainability* in the regression - if and how the relative weights of the partial ratings shifted after this event. We cannot directly address this question due to the reduction in sample size it would imply. However, using the Chow test[23], we can check if the coefficients in column (6) are the same before and after 2006, or whether the revision of the evaluation criteria marks a structural change between these two time periods.[24] We do not find evidence for a structural break (results are available from the authors on request).[25] This result is plausible given that the alignment with OECD/DAC criteria in 2006 did not revise the former evaluation criteria of KfW, but rather extended them by the dimension *sustainability* as a stand-alone criterion. Moreover, the quantitative and qualitative indicators KfW defined to assess the impact of drinking water aid projects neither changed after 2006.

Table 3 shows the correlation between the evaluation ratings and the project indicators. We expect that the project indicators are highly and positively correlated with the ratings. We further expect the correlation to be stronger when using the *effectiveness* evaluation rating, since it is explicitly linked to the chosen indicators of improved water supply. We combine the water and population related indicators together as explanatory variables in one regression, since evaluators have to consider them simultaneously when evaluating a project. The variables are paired as follows. The regressions in columns (1) and (4) correspond to the total supply levels at t1 (after project completion); the regressions in columns (2) and (5) correspond to the change in supply

[23] For a detailed description of the Chow test, see Wooldrige Chapter 7 and 13.

[24] The period before 2006 comprises 77.33 percent of the observations in the estimation sample, the period after 2006 correspondingly comprises 22.67 of the observations.

[25] The value of the F-statistic (1.997) is smaller than the critical value (3.268). Hence we cannot reject the null hypothesis of "no structural break".

levels between t0 (before the project) and t1 (after the project); and the regressions in columns (3) and (6) correspond to the project goal achievement at t1. The dependent variable for the regression in columns (1)-(3) is the *overall* evaluation rating, for columns (4)-(6) it is the *effectiveness* rating. Collinearity between the independent variable pairs is not an issue (see correlation matrix in Table A3 in the Appendix). We also add the squared term of *daily per capita safe water consumption* to the model presented in column (1) and (4) to control for non-linear effects of water consumption per day on project ratings. We also included the squared term of *number of people with access to an improved water source* in previous specifications but it turned out to be insignificant.

Table 3 Correlation between evaluation ratings and project indicators

	Overall Evaluation Rating			Effectiveness Evaluation Rating		
	(1)	(2)	(3)	(4)	(5)	(6)
	Total water supply at t1					
Water pcl	**1.250***			**1.404***		
	(0.063)			(0.097)		
Water pcl^2	**-0.186****			**-0.199***		
	(0.035)			(0.082)		
Population	**0.227*****			**0.293*****		
	(0.000)			(0.000)		
	Change in water supply between t0 and t1					
Change Water pcl		-0.033			-0.189	
		(0.891)			(0.558)	
Change Population		**0.679****			**1.113*****	
		(0.031)			(0.001)	
	Goal achievement water supply at t1					
Goal Achievement Water pcl			-0.097			-0.100
			(0.547)			(0.639)
Goal Achievement Population			**0.222****			**0.200***
			(0.035)			(0.096)
Constant	-0.543	3.636***	3.857***	-1.595	3.746***	4.016***
	(0.708)	(0.000)	(0.000)	(0.370)	(0.000)	(0.000)
Observations	135	78	130	88	54	85
R2	0.130	0.058	0.038	0.162	0.174	0.036
Adjusted R2	0.110	0.033	0.023	0.132	0.141	0.013

Notes: Standard errors in parentheses, *significant at the 10% level, ** significant at 5%; *** significant at 1%; see Table 1 for precise definitions of dependent variables.

The results show that the project indicators affect the ratings differently: *per capita safe water consumption* increases the chances of a good evaluation rating, but only up to a certain level, while a higher *number of people with access to a safe water source* always increases evaluation ratings (Table 3, column 1 and 4). A plausible explanation for diminishing marginal returns in *per capita safe water consumption* is that KfW defines "adequate" consumption level standards for the target group. Depending on the project area and the water supply system in question, these standards range between a minimum of 20 liters and a maximum of 180 liters per capita and day. Exceeding the upper bound has also happened in projects.

Most interestingly, the change in supply levels (column (2) and (5)), or the extent to which targets were achieved (column (3) and (6)), has a smaller impact on evaluation ratings than absolute supply levels. Moreover, only the population related variables are correlated with the evaluation ratings whereas the water consumption related variables do not have any effect on evaluation ratings. Last, and as expected the adjusted R2 of effectiveness ratings are somewhat larger than the adjusted R2 of overall ratings, but unexpectedly R2 for the effectiveness ratings are also low, even though the effectiveness ratings should directly evaluate improvements in water supply (access, quantity, and also quality, on which information is however barely available).

Next, we use both evaluation ratings and water supply indicators as dependent variables (see Table 4) and examine (a) if and how country-level characteristics affect the performance of drinking water projects and (b) how the answer to this question varies with the way we measure project performance. The R2 values of the regressions presented in Table 4 show that overall country-level characteristics explain little of the variation in performance. This result parallels the findings of other studies, which show that country-level characteristics explain only a small share of the variation in the World Bank performance ratings (see e.g. Denizer et al., 2013; Guillaumont and Laajaj, 2006). Noteworthy exceptions to this pattern are the R2s of the regressions in which we accounted for the baseline value (at project start) for *water consumption per capita* (column (3)) and *number of people with access to an improved water source* (column (6)). The coefficients of both variables in the respective regressions are highly significant and indicate that the level of *per capita water consumption* and the *total number of people with access to an improved water source* at the time of evaluation mainly depends on the values of these variables at project start. This is also true for the regressions in Table 5 and Table 6 (see columns (3) and (6)).

According to the results in column (1) and (2) the event of a conflict (taking place at any point in time in the period between appraisal and evaluation) significantly negatively affects the evaluation ratings. Interestingly, *conflict* does not seem to affect project performance when it is measured in terms of (changes in) water supply indicators (Table 4, columns 3-8). Similarly, a better political environment positively influences evaluation ratings whereas the evidence of its impact on water supply is rather mixed. Another interesting finding is that increasing levels of bilateral aid per capita have a negative impact on the evaluation ratings (column 1 and 2).[26] This result is akin to Guillaumont and Laajaj's (2006) finding that increasing levels of ODA, measured in percent of GDP, reduce a project's success probability. However, in our case, the result is only relevant for the evaluation ratings (which are similar to the outcome variables applied in previous literature). There is no clear relationship between the average flow of bilateral aid to a recipient country (between t0 and t1) and project indicators. Last, we observe that a higher GDP per capita is positively correlated with (changes in) water consumption (see

[26] Aid squared turned out to be insignificant and was therefore excluded for the final specification.

Table 4, column 3 and 4). In contrast, GDP per capita has no effect on evaluation ratings and a rather negative effect on the change in population with access to improved water. A possible explanation for this result is that countries with higher GDP per capita often have relatively high access rates at project start.

Table 4 Correlation between project performance and country characteristics

	Evaluation Ratings		Water Consumption per Capita			Number of People with Access		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Overall	Effective-ness	Water pcl	Change water pcl	Achievement water pcl	Population	Change population	Achievement population
Avg. GDP per capita [a)]	-0.182	-0.158	**0.187***	**0.166***	-0.032	0.184	**-0.107***	0.018
	(0.192)	(0.406)	**(0.057)**	**(0.056)**	(0.666)	(0.361)	(0.085)	(0.891)
Avg. Aid per capita	**-0.063****	**-0.067****	0.029	0.003	0.011	-0.099**	0.000	-0.016
	(0.016)	**(0.035)**	(0.125)	(0.868)	(0.405)	(0.016)	(0.974)	(0.521)
Avg. Inflation Rate	-0.001**	-0.001	0.001	-0.000	-0.000	-0.000	0.000	0.000
	(0.030)	(0.473)	(0.450)	(0.783)	(0.388)	(0.665)	(0.551)	(0.797)
Polity Index	**0.028***	**0.045****	**-0.019***	**-0.018***	**0.013***	0.003	-0.002	**0.025***
	(0.065)	**(0.029)**	**(0.092)**	**(0.063)**	**(0.092)**	(0.912)	(0.773)	**(0.083)**
Conflict	**-0.229***	**-0.299***	-0.049	-0.047	-0.024	0.092	0.023	**0.233***
	(0.089)	**(0.084)**	(0.613)	(0.581)	(0.727)	(0.639)	(0.705)	**(0.059)**
Africa Region	-0.100	-0.062	**-0.440****	0.166	**-0.227***	0.028	0.090	0.029
	(0.692)	(0.855)	**(0.022)**	(0.299)	**(0.086)**	(0.937)	(0.416)	(0.900)
MENA Region	0.423	**0.938****	-0.021	0.090	-0.253	0.427	0.136	0.276
	(0.196)	**(0.029)**	(0.921)	(0.635)	(0.128)	(0.383)	(0.369)	(0.370)
Water pcl at Start			**0.350*****					
			(0.000)					
Population at Start						**0.206*****		
						(0.000)		
Constant	5.608***	5.552***	1.659**	-1.031*	1.283**	8.807***	1.126**	0.806
	(0.000)	(0.000)	(0.030)	(0.097)	(0.015)	(0.000)	(0.012)	(0.383)
R2	0.115	0.155	0.634	0.085	0.114	0.432	0.134	0.071
Adjusted R2	0.066	0.075	0.594	0.000	0.060	0.374	0.057	0.014
N	133	82	83	83	122	87	87	121
Notes: p-values in parentheses, *significant at 10% level, ** significant at 5%; *** significant at 1%; [a)] logarithm								

Tables 5 and 6 focus on the questions (a) how much of the variation in the different performance indicators is explained by project-level variables, and (b) how does the impact of these variables on project performance vary with the indicator of project performance chosen. To this end, we begin with an analysis of the relationship between *project management* characteristics and the various performance indicators (Table 5). Then, we examine the relationship between specific *project design* variables and the various performance indicators (Table 6).

Table 5 Correlation between project performance and project management

	Ratings		Water Consumption per Capita			Number of People with Access		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Overall	Effectiveness	Water pcl	Change water pcl	Achievement water pcl	Population	Change population	Achievement population
Months Past Project	0.002	0.002	-0.000	-0.002	-0.002	0.001	-0.001	0.001
	(0.636)	(0.580)	(0.987)	(0.221)	(0.288)	(0.774)	(0.454)	(0.422)
Project Duration	**-0.010****	-0.003	**0.004***	0.002	0.001	0.005	0.002	**0.003****
	(0.017)	(0.492)	(0.061)	(0.326)	(0.590)	(0.245)	(0.136)	(0.042)
Second Phase [a)]	0.053	0.116	0.083	-0.037	0.158	-0.181	-0.020	-0.063
	(0.857)	(0.705)	(0.622)	(0.788)	(0.212)	(0.548)	(0.855)	(0.538)
Third Phase	0.493	-0.384	**1.202*****	**0.741****	**0.626***	0.174	0.441	-0.203
	(0.500)	(0.614)	(0.003)	(0.021)	(0.088)	(0.872)	(0.240)	(0.481)
Risk at Project Start	**-0.564*****	**-0.532****	-0.043	0.009	**-0.160***	-0.172	-0.027	-0.084
	(0.007)	(0.015)	(0.697)	(0.917)	(0.054)	(0.417)	(0.724)	(0.203)
Total Project Costs USD	**0.416*****	**0.364****	0.002	-0.044	-0.011	**0.565*****	-0.014	0.059
	(0.002)	(0.011)	(0.979)	(0.478)	(0.844)	(0.000)	(0.779)	(0.172)
Recipient Share	-0.647	-0.163	**0.634****	-0.152	-0.238	0.362	-0.090	-0.121
	(0.246)	(0.779)	(0.031)	(0.510)	(0.307)	(0.466)	(0.623)	(0.510)
Water pcl at Start			**0.438*****					
			(0.000)					
Population at Start						0.135***		
						(0.000)		
Constant	-1.220	-0.657	2.175*	0.951	1.551*	1.337	0.766	0.051
	(0.583)	(0.776)	(0.075)	(0.350)	(0.085)	(0.561)	(0.351)	(0.941)
R2	0.192	0.156	0.625	0.131	0.099	0.491	0.076	0.092
Adjusted R2	0.121	0.082	0.581	0.043	0.041	0.434	-0.014	0.032
N	87	87	77	77	116	80	80	115

Notes: p-values in parentheses, *significant at 10% level, ** significant at 5%; *** significant at 1%; [a)] phase 1 is the left out category

Table 5, columns 1 and 2, shows that the duration of a project's implementation has a negative impact on project evaluations. This result is in line with Denizer et al. (2013) who find this effect for World Bank projects. However, when we use the water supply indicators as dependent variables, we do not find evidence that a longer implementation period leads to lower project performance. Our analysis on evaluation ratings (columns 1 and 2) further supports another result found in previous studies: a high perceived project failure risk at t0 has a strong negative effect on evaluation ratings. Similar to the variable project duration, high project failure risk turns insignificant once we use water supply indicators as a measure of project performance. Our interpretation of this result is that evaluators are more likely to give lower evaluation ratings if they know that the risk of a project was perceived to be high at project start. Last, and again in line with previous research on the topic, project costs have a positive impact on evaluation ratings (Table 5, columns 1 and 2). Kilby (2000) suggests that this might be either because grants or loans in "well established sub-sectors" are larger or because committed borrowing governments receive larger loans or grants. Again this variable has no effect on improved water access, except for the number of people provided with improved water access.

The only project management variable that is strongly (positively) correlated with improvements in water access is the 3rd phase of a project (Table 5, column 3-5). The fact that a project continued past the initial phase may signal the German Government´s general interest for a longer-term engagement in the recipient country's drinking water sector. It is furthermore possible that later phases are more successful, because of learning effects.

Table 6 Correlation between project performance and project design

	Evaluation Ratings		Water Consumption per Capita			Number of People with Access		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Overall	Effective-ness	Water pcl	Change water pcl	Achievement water pcl	Population	Change population	Achievement population
Tariff per cbm used [a)]	**0.963****	**0.840***	-0.295	-0.061	-0.162	-0.259	**-0.402****	-0.202
	(0.038)	**(0.083)**	(0.275)	(0.785)	(0.403)	(0.675)	**(0.026)**	(0.146)
Mixed System [b)]	-1.316*	-1.353*	-0.643	0.057	-0.203	-0.554	0.140	-0.243
	(0.084)	(0.090)	(0.236)	(0.901)	(0.487)	(0.574)	(0.628)	(0.242)
Piped System	**-0.952***	**-1.140****	**1.270*****	**0.747*****	0.020	-0.445	0.176	**-0.257***
	(0.068)	**(0.037)**	**(0.000)**	**(0.001)**	(0.924)	(0.466)	(0.312)	(0.089)
Urban	0.451	0.385	-0.443	-0.158	-0.111	-0.152	-0.175	-0.016
	(0.276)	(0.374)	(0.118)	(0.494)	(0.522)	(0.754)	(0.208)	(0.897)
User Committee	0.579	0.361	0.237	**0.499*****	0.197	0.509	0.002	0.127
	(0.230)	(0.475)	(0.291)	**(0.007)**	(0.267)	(0.334)	(0.988)	(0.339)
Wastewater Component	0.183	0.218	**0.283****	-0.129	**0.195***	-0.259	-0.013	0.108
	(0.458)	(0.398)	**(0.023)**	(0.172)	**(0.058)**	(0.315)	(0.860)	(0.145)
Hygiene Promotion	0.157	0.034	**0.227***	**0.273*****	-0.073	-0.086	**0.266*****	-0.112
	(0.588)	(0.911)	**(0.094)**	**(0.006)**	(0.527)	(0.764)	**(0.001)**	(0.187)
Institutional Support	0.367	0.291	**0.222***	0.021	0.054	-0.020	**-0.130***	0.038
	(0.231)	(0.365)	**(0.058)**	(0.828)	(0.647)	(0.939)	**(0.094)**	(0.655)
Low Water Resources	-0.300	-0.353	0.078	0.048	0.106	-0.384	-0.096	0.073
	(0.261)	(0.205)	(0.492)	(0.611)	(0.306)	(0.124)	(0.191)	(0.336)
Water pcl at Start			**0.296*****					
			(0.000)					
Population at Start						**0.235*****		
						(0.000)		
Constant	3.958***	4.449***	2.146***	-0.502**	0.793***	10.39***	0.618***	1.188***
	(0.000)	(0.000)	(0.000)	(0.024)	(0.001)	(0.000)	(0.001)	(0.000)
R2	0.126	0.100	0.673	0.308	0.070	0.463	0.307	0.113
Adjusted R2	0.017	0.012	0.631	0.229	0.003	0.394	0.228	0.043
N	82	82	89	89	125	89	89	124

Notes: p-values in parentheses, *significant at 10% level, ** significant at 5%; *** significant at 1%; [a)] logarithm; [b)] left-out category is non-piped systems.

Columns 1 and 2 in Table 6 show that the tariff rate per cubic meter of consumed water significantly increases evaluation ratings. The fact that a higher tariff is positively and significantly related to a higher overall project performance rating is coherent with one of the key propositions made in the impact chain of water supply (see also Section 3.2): cost coverage through tariffs is considered important for sustainable drinking water provision. Anecdotal evidence from project reports, which frequently point out that tariffs need to be sufficiently high in order to guarantee a sustainable effect from the project, further supports this finding.[27] These findings are at odds with the estimates in column 3-8 of Table 6 (and Table A4 in the Appendix), which indicate that a higher tariff is rather negatively correlated with per capita water consumption levels and the population reached. This result describes an evident relationship and,

[27] An evaluation report for an urban drinking water project in Vietnam for example explicitly lists full cost coverage as an indicator to measure the efficiency and sustainability of the project goal (safe drinking water supply). However, the report also points out that full cost coverage is not an appropriate indicator to measure the actual achievement of the supply goals.

there is, again, anecdotal evidence from project reports where evaluators discuss the problem that high tariffs can have a negative effect on consumption levels.[28]

A further interesting finding is that working with piped systems apparently reduces the probability of a good evaluation rating, as can be seen in column 1 and 2 of Table 6. Piped systems in developing countries, especially those which need rehabilitation within the scope of aid projects, are often characterized by high unaccounted water losses which negatively affect the sustainability of a project and therefore overall evaluation rating. Again, the results change when we look at the project indicators (Table 6, column 3-8). When project performance is measured in terms of total per capita water consumption at evaluation, the presence of a piped system is positively correlated with higher consumption levels: first, purely piped systems usually have larger capacities than non-piped systems, and second, micro-level research has shown that water consumption levels increase significantly with (privately) piped water systems (see e.g. Devoto, 2013; WHO, 2003).

Last, we find that hygiene promotion has a significant positive impact on the change in consumption levels between project start and evaluation and that waste water treatment significantly increases water consumption levels. For waste water this is a straightforward relationship, for hygiene promotion this is an encouraging result: hygiene promotion activities seem to be successful in sensitizing target groups to the benefits of safe drinking water consumption.

Overall, we find that any conclusion on the determinants of projects' aid effectiveness highly depends on the performance indicator chosen, and that the differences are most pronounced between evaluation (*overall and effectiveness*) ratings on the one hand and water supply indicators on the other. Interestingly, we also find (see Tables 3-6, columns 1 and 2) that there is apparently no conceptual difference between the overall and the effectiveness rating. The estimated coefficients are very similar regardless of whether we use the overall or effectiveness evaluation rating. This implies that these ratings are (almost) interchangeable.

Moreover, we find that evaluation ratings are mostly affected by project management whereas water access and consumption is best explained by project design. This is also confirmed by Table 7, which compares the adjusted R2 of the various specifications. The last row of Table 7 shows the variance in project performance explained by the combination of all country, management, and design variables. The corresponding regression results can be found in Table A4 in the Appendix. Note that the coefficients reported in Table 4-6 do not change qualitatively once we control for all variables, only the significance levels decrease somewhat for some

[28] An evaluation of a drinking water project in a semi-urban area in Burkina Faso states that it is likely that the increased tariff rate led to a reduction in consumption levels, especially of poorer population segments. At the same time, the report emphasizes the positive effect of the higher tariff rate, namely improved cost coverage of the drinking water provision.

explanatory variables, which is expected given the rather small sample size. The explained variation in project performance increases considerably when we take account of all country and project-level variables. However, the share of variation in project outcomes we are able to explain still remains moderate.

Table 7 Variance in performance explained by country, project management and project design

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Ratings		Water Consumption per Capita			Number of People with Access		
	Overall	Effective-ness	Water pcl	Change water pcl	Achievement water pcl	Popu-lation	Change population	Achievement population
Country Characteristics	0.066	0.075	0.594	0.000	**0.060**	0.374	0.057	0.014
Project Management	**0.121**	**0.082**	0.581	0.043	0.041	**0.434**	-0.014	0.032
Project Design	0.017	0.012	**0.631**	**0.229**	0.003	0.394	**0.228**	**0.043**
Country, Management, and Design	0.190	0.103	0.703	0.301	0.055	0.541	0.231	0.143

Notes: Values in the first 3 rows are the Adjusted R2 values from Table 4-6. The values in the last row are the Adjusted R2 values from Table A4.

5. Conclusion

We analyzed the correlates of project performance for 150 water supply projects by comparing evaluation ratings with sector-specific and internationally agreed upon indicators. First, we find that the correlation between the overall rating and the effectiveness rating is positive and (very) high whereas the correlation between (overall and effectiveness) evaluation ratings and the indicators measuring improved water access is positive but weak. Second, our analysis shows that the explanatory variables affect project performance differently depending on how project performance is measured and that this leads to different, and even contradictory, conclusions about what increases aid effectiveness. For example, higher tariffs have a positive impact on evaluation ratings but a negative impact on water supply indicators; proxies for political and economic stability have a negative impact on evaluation ratings but are rather weakly correlated with the indicators of improved water supply. Third, we find that project management variables have a higher impact on evaluation ratings whereas project design variables have a higher impact on the indicators measuring improved access to safe drinking water. Last, the explanatory power of our models increases considerably when we take into account project level variables (in addition to general country characteristics).

In the broader context of the aid effectiveness debate this study has the following implications. It is important to take into account aid heterogeneity. Project aid data should be analyzed at the sector level with meaningful definitions and measurement units for outcome indicators that allow for an in-depth assessment of projects' aid effectiveness. A similar argument has been made in other studies (Clemens, Radelet and Bhavani, 2004; Dreher, Nunnenkamp and Thiele, 2008; Findley et al., 2010; Mavrotas, 2005; Ndikumana, 2012). This would also allow for the comparison of observational project-level studies to experimental micro-level studies to test whether insights from RCTs might be externally valid across projects and countries.

Hence, it is important that donors systematically collect and apply objectively comparable project performance indicators, at least in sectors for which such an approach is feasible, and that these indicators meet, where possible, internationally agreed standards. The latter is especially important when donor organizations have a general interest in helping to close the "micro-macro gap" in the aid effectiveness debate. Specifically the harmonization of donor efforts in recipient countries (to further the implementation of the Paris Declaration and Accra Agenda for Action[29]) requires the definition of joint project goals and joint indicators to evaluate the achievement of these goals.

Last, our analysis has shown that evaluation ratings rather seem to reflect the donor's concept of what constitutes project performance and how projects should be evaluated. Hence, when the objective is to assess if a project was effective in terms of achieving its direct goals (and to find out which factors helped achieving these goals), evaluation ratings should not be used. The advantage of evaluation ratings is that they go beyond mere water supply targets (in our case) to reflect project success. They are however donor specific. The advantage of supply indicators is that they are not donor specific, but are internationally agreed upon drinking water indicators stipulated in the Millennium Development Goals (MDGs) and the United Nations Declaration of the Human Right to Water.

[29] For further information see: http://www.oecd.org/dac/effectiveness/parisdeclarationandaccraagendaforaction.htm

Appendix

Table A1: Country coverage of KfW water aid data set

East Asia & Pacific	Europe & Central Asia	Latin America & Caribbean	Middle East & North Africa	South Asia	Sub-Saharan Africa
China (4)	Albania (6)	Bolivia (4)	Djibouti (1)	India (2)	Benin (2)
Indonesia (5)	Azerbaijan (1)	Brazil (5)	Egypt (1)	Maldives (1)	Botswana (2)
Lao PDR (1)	Georgia (1)	Costa Rica (1)	Jordan (1)	Pakistan (1)	Burkina Faso (4)
Samoa (1)	Kosovo (2)	Ecuador (4)	Morocco (6)		Cameroon (1)
Vietnam (1)	Turkey (4)	El Salvador (1)	Tunisia (2)		Cape Verde (1)
	Uzbekistan (1)	Guatemala (1)	West Bank/Gaza (2)		CAR (1)
		Honduras (1)	Yemen, Rep. (2)		Chad (2)
		Nicaragua (1)			Congo, Dem. Rep. (2)
		Paraguay (1)			Congo, Rep. (1)
		Peru (4)			Côte d'Ivoire (1)
					Ethiopia (3)
					Ghana (4)
					Guinea (1)
					Kenya (5)
					Lesotho (3)
					Madagascar (1)
					Malawi (2)
					Mali (9)
					Mauritania (1)
					Mozambique (1)
					Namibia (3)
					Niger (5)
					Nigeria (1)
					Rwanda (2)
					Senegal (6)
					Sudan (1)
					Tanzania (8)
					Togo (1)
					Uganda (1)
					Zambia (5)
					Zimbabwe (1)
(12)	(15)	(23)	(15)	(4)	(81)

Notes: Number of projects in parentheses.

Table A2: Summary statistics

	N	mean	min	max
KfW project ratings				
Overall Rating	150	3.98	2	6
Effectiveness Rating	93	4.14	2	6
Efficiency Rating	93	3.83	2	6
Sustainability Rating	35	4.06	3	6
Impact Rating	93	4.34	2	6
Relevance Rating	93	4.6	2	6
Water consumption p.c./day in liters				
Water pcl t0_baseline	93	57.95	0.00	226.00[30]
Water pcl t1_actual	139	66.58	4.00	300.00
Water pcl t1_target	141	74.74	9.50	250.00
% Change Water Consumption	93	0.17	-0.98	1.00
Goal Achievement Water Consumption	137	0.94	0.20	3.60
Population with safe water access				
Population With Access t0_baseline	99	243'898	0.000	6'700'000
Population With Access t1_actual	139	401'551	2'364	16'000'000
Population With Access t1_target	143	401'994	8'000	10'000'000
% Change Population With Safe Water Access	96	0.499	-1.571	1.000
Goal Achievement Population With Access	136	0.993	0.063	9.097
Country characteristics				
Average GDP per capita	144	1'018.520	141.950	5'151.698
Average Bilateral Aid per capita	144	3.941	-0.289	21.304
Average Inflation	141	74.785	1.483	2'459.968[31]
Polity2 Score at Project Start	141	-1.348	-9	10
Conflict	150	0.61	0	2
Sub-Saharan Africa	150	0.54	0	1
Middle East/North Arica	150	0.20	0	1
Project management variables				
Months Past Project Closure	150	48.080	0	144.000
Project Duration in Months	150	70.03	17.5	216
Risk at Project Start	129	2.271	1	4
Total Project Cost (USD)	142	18'435'169	1'656'902	184'779'632
Recipient Share in Project Cost %	142	0.197	0.0	0.89
First Project Phase	150	0.246	0	1
Project design variables				
Tariff per cbm of water Consumed at t1 (USD)	135	0.46	0.000	5.52
Piped System	149	0.83	0	1
Urban	150	0.726	0	1
User Committee	149	0.275	0	1
Wastewater Component	150	0.346	0	1
Hygiene Promotion	150	0.446	0	1
Institutional Support	149	0.241	0	1
Low Water Resources	149	0.82	0	1

[30] We excluded one project in Samoa, Turkey, Albania, and Egypt from the summary statistics for *water pcl t0_baseline*, *water pcl t1_actual*, *water pcl t1_target*. These projects are outliers since their goal was to reduce "excessive" water consumption.
[31] The summary statistic for this variable includes very high inflation values for Congo, Dem. Rep. The median is at 7.51.

Table A3: Correlation matrix of project indicator variables

	water pcl	population t1	change water pcl	change population	achievement water pcl	achievement population
Water pcl t1	1					
Population t1	**0.283*****	1				
Change Water pcl	-0.0947	-0.0814	1			
Change Population	-0.377***	-0.0199	**0.131**	1		
Achievement Water pcl	0.430***	-0.0373	0.396***	-0.152	1	
Achievement Population	0.0871	0.250**	-0.0861	0.208*	**-0.0672**	1

Notes: p-values in parentheses, $^{+}$ significant at 15 percent level, ***significant at 1 percent level, **significant at 5 percent level; see Table 1 for precise definitions of dependent variables.

Table A4: Correlation between project performance and country, management & design

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Ratings		**Water Consumption pcl**			**Number of people with access**		
	Overall	**Effectiveness**	**Water**	**Change**	**Achievement**	**Population**	**Change**	**Achievement**
	Country Characteristics							
Avg. GDP per capita[a)]	**-0.674*****	**-0.630****	0.121	**0.255****	0.032	-0.229	**-0.216****	**-0.214****
	(0.010)	(0.030)	(0.313)	(0.010)	(0.744)	(0.356)	(0.015)	(0.011)
Avg. Aid per capita	-0.061	-0.130	**0.175***	0.092	-0.040	-0.113	0.026	0.011
	(0.712)	(0.482)	(0.056)	(0.240)	(0.521)	(0.535)	(0.692)	(0.834)
Avg. Inflation Rate	-0.000	-0.002	0.001	0.000	0.000	0.002	-0.000	0.000
	(0.794)	(0.411)	(0.374)	(0.903)	(0.649)	(0.284)	(0.623)	(0.684)
Polity Index	0.031	**0.046***	-0.005	**-0.020***	**0.017***	-0.013	-0.000	**0.012**$^{+}$
	(0.180)	(0.072)	(0.702)	(0.072)	(0.077)	(0.605)	(0.966)	(0.148)
Conflict	-0.294	-0.250	0.024	0.027	0.020	-0.079	0.047	-0.042
	(0.233)	(0.362)	(0.831)	(0.784)	(0.817)	(0.720)	(0.552)	(0.578)
Africa Region	-0.470	-0.581	-0.229	0.017	-0.052	-0.069	-0.020	-0.190
	(0.292)	(0.240)	(0.289)	(0.926)	(0.769)	(0.868)	(0.894)	(0.207)
MENA Region	0.710	**0.869**$^{+}$	-0.213	0.220	-0.200	0.343	0.160	0.143
	(0.183)	(0.142)	(0.479)	(0.382)	(0.377)	(0.584)	(0.477)	(0.463)
	Project Management							
Months Past Project	0.001	0.002	-0.003	-0.003	-0.002	0.003	-0.001	0.001
	(0.894)	(0.659)	(0.175)	(0.207)	(0.270)	(0.550)	(0.531)	(0.584)
Project Duration	-0.006	0.000	0.003	0.002	-0.002	0.006	**0.003***	**0.004****
	(0.213)	(0.972)	(0.241)	(0.429)	(0.353)	(0.177)	(0.093)	(0.037)
Second Phase[b)]	0.010	0.050	0.136	-0.010	0.163	0.049	-0.169	-0.073
	(0.979)	(0.906)	(0.445)	(0.948)	(0.285)	(0.889)	(0.173)	(0.576)
Third Phase	0.616	-0.076	0.109	**-0.755***	0.333	0.244	**0.896***	-0.367
	(0.512)	(0.942)	(0.836)	(0.094)	(0.407)	(0.866)	(0.060)	(0.286)
Risk at Project Start	**-0.618****	**-0.654***	0.023	-0.022	-0.020	**-0.441**$^{+}$	**-0.190***	**-0.160***
	(0.049)	(0.061)	(0.874)	(0.853)	(0.852)	(0.110)	(0.054)	(0.086)
Total Project Cost USD	**0.493*****	**0.308**$^{+}$	0.004	**-0.109**$^{+}$	0.044	**0.659*****	0.054	**0.095**$^{+}$
	(0.009)	(0.142)	(0.966)	(0.142)	(0.511)	(0.000)	(0.347)	(0.100)
Recipient Share	0.021	0.432	0.442	-0.170	-0.210	0.598	0.122	0.045
	(0.978)	(0.606)	(0.218)	(0.568)	(0.472)	(0.383)	(0.621)	(0.859)
	Project Design							
Tariff per cbm USD[a)]	0.677	0.778	**-0.629**$^{+}$	**-0.555***	**-0.456***	-0.389	-0.308	-0.234
	(0.359)	(0.343)	(0.106)	(0.096)	(0.094)	(0.631)	(0.289)	(0.316)
Mixed System[c)]	-0.634	-0.831	-0.461	0.141	-0.067	0.445	0.345	-0.278
	(0.455)	(0.377)	(0.435)	(0.780)	(0.816)	(0.651)	(0.328)	(0.258)
Piped System	-0.804	**-1.142**$^{+}$	**1.175*****	**0.953*****	0.319	-1.003	-0.148	-0.223
	(0.236)	(0.129)	(0.003)	(0.004)	(0.180)	(0.184)	(0.581)	(0.278)
Urban	0.008	0.061	-0.069	0.116	-0.013	0.177	-0.072	-0.006
	(0.977)	(0.833)	(0.671)	(0.400)	(0.895)	(0.472)	(0.414)	(0.942)
User committee	0.076	-0.104	0.426	**1.088*****	**0.297**$^{+}$	0.115	**-0.396***	0.121
	(0.900)	(0.876)	(0.196)	(0.000)	(0.128)	(0.862)	(0.092)	(0.481)
Wasterwater Component	-0.130	-0.087	0.131	0.033	0.083	**-0.562****	0.015	0.038
	(0.643)	(0.780)	(0.400)	(0.801)	(0.463)	(0.038)	(0.878)	(0.692)
Hygiene Promotion	0.029	0.026	0.015	**0.474*****	0.011	0.024	0.138	0.022
	(0.933)	(0.946)	(0.929)	(0.000)	(0.933)	(0.936)	(0.196)	(0.844)
Institutional Support	0.037	-0.130	**0.333***	0.089	0.111	-0.103	**-0.159**$^{+}$	0.122
	(0.919)	(0.750)	(0.055)	(0.553)	(0.382)	(0.730)	(0.123)	(0.268)
Low Water Resources	-0.145	-0.094	0.151	-0.013	0.214	-0.070	-0.118	0.142
	(0.685)	(0.813)	(0.362)	(0.927)	(0.106)	(0.819)	(0.280)	(0.222)
Water pcl at Project Start			**0.304*****					
			(0.000)					
Population at Project Start						**0.145*****		
						(0.001)		

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Ratings		**Water Consumption pcl**			**Number of people with access**		
	Overall	**Effectiveness**	**Water**	**Change**	**Achievement**	**Population**	**Change**	**Achievement**
Constant	2.711	5.626	0.924	-0.901	-0.039	2.638	1.756*	1.235
	(0.431)	(0.141)	(0.562)	(0.508)	(0.975)	(0.390)	(0.098)	(0.247)
R2	0.468	0.411	0.813	0.548	0.286	0.706	0.495	0.357
Adjusted R2	0.190	0.103	0.703	0.301	0.055	0.541	0.231	0.143
N	68	68	66	66	95	68	68	93

Notes: p-values in parentheses, [+] significant at 15 percent level, ***significant at 1 percent level, **significant at 5 percent level, *significant at 10 percent level; [a)] logarithm; [b)] left-out category is phase 1; [c)] left-out category is non-piped systems.

References

ANAND P.P., 2013. “Every drop counts. Assessing aid for water and sanitation,” *UNU-WIDER Working Paper*, 2013/025, United Nations University, Helsinki.

BUDDS, J., McGRANAHAN, G., 2003. “Are the debates on water privatization missing the point? Evidence from Africa, Asia and Latin America,” *Environment and Urbanization,* 15(2), 87-114.

BURNSIDE, C., DOLLAR, D., 2000. “Aid, Policies and Growth,” *American Economic Review*, 90(4), 847-868.

BURNSIDE, C., DOLLAR, D., 2004. “Aid, Policies and Growth: Revisiting the Evidence,” *World Bank Policy Research Working Paper,* 3251, World Bank, Washington, D.C.

CLEMENS, M., RADELET, S., BHAVNANI, R., 2004. “Counting chickens when they hatch: The short-term effect of aid on growth”, Working Paper No. 44 (July 2004), Center for Global Development, Washington, D.C.

COLLIER, P., DEHN, J., 2001. “Aid, Shocks, and Growth,” *World Bank Policy Research Working Paper,* 2688, World Bank, Washington, D.C.

COLLIER P., DOLLAR, D., 2002. “Aid Allocation and Poverty Reduction,” *European Economic Review*, 46, 1475-1500.

DALGAARD, C.-J., HANSEN, H., TARP, F., 2004. ”On the Empirics of Foreign Aid and Growth,” *The Economic Journal*, 114, 191-216.

DENIZER, C., KAUFMANN, D., KRAAY, A., 2013. “Good Countries or Good Projects? Macro and Micro Correlates of World Bank Project Performance,” *Journal of Development Economics,* in press.

DEVOTO, F., DUFLOS, E., DUPAS, P., PARIENTE, W., PONS, V., 2013. “Happiness on tap: Piped water adoption in urban Morocco,” *American Economic Journal: Economic Policy,* in press.

DOLLAR, D., LEVINE, V., 2005. “Sowing and Reaping: Institutional Quality and Project Outcomes in Developing Countries,” *World Bank Policy Research Working Paper*, 3524, World Bank, Washington D.C.

DOLLAR, D., SVENSSON, J., 2000. “What Explains the Success and Failure of Structural Adjustment Programs?” *The Economic Journal*, 110(446), 894-917.

DREHER, A., KLASEN, S., VREELAND, J., WERKER, E., 2013. “The Costs of Favoritism: Is Politically-Driven Aid Less Effective?” *Economic Development and Cultural Change,* in press.

DREHER, A., NUNNENKAMP, R., THIEL, S., THIELE, R., 2008. “Does Aid for Education Educate Children? Evidence from Panel Data” *The World Bank Economic Review*, 22(2), 291-314.

GUILLAUMONT, P., LAAJAJ, R., 2006. “When instability increases the effectiveness of aid projects,” *World Bank Policy Research Working Paper*, 4034, World Bank, Washington D.C.

HEMMER, H., LORENZ, A., 2003. "What determines the success or failure of German Bilateral Financial Aid?" *Review of World Economics*, 139(3), 507-549.

ISHAM, J., KAUFMANN, D., PRITCHETT, L., 1997. "Civil Liberties, Democracy, and the Performance of Government Projects," *World Bank Economic Review*, 11(2), 219-242.

ISHAM, J., KAUFMANN, D., 1999. "The Forgotten Rationale For Policy Reform: The Productivity of Investment Projects," *Quarterly Journal of Economics*, 114(1), 149-84.

IVANOVA, A., MAYER, W., MOURMOURAS, A., ANAYIOTOS, G., 2001. "What determines the implementation of IMF-Supported programs?" *IMF Working Papers*, W/03/8, International Monetary Fund, Washington D.C.

KILBY, C., 2000. "Supervision and Performance: The case of World Bank Projects," *Journal of Development Economics*, 62(1), 233-259.

KfW, 2004a. "Woran scheitern so viele Vorhaben der Siedlungswasserwirtschaft?" unpublished, KfW Entwicklungsbank, Frankfurt a.M.

KfW, 2004b. "Evaluierung in der Siedlungswasserwirtschaft," unpublished, KfW Entwicklungsbank, Frankfurt a.M.

KfW, 2005. "Wasserversorgung Bengkulu, Indonesien," unpublished, KfW Entwicklungsbank, Frankfurt a.M.

KfW, 2006a. "Betriebswirtschaftliche Projektrechnung für die Siedlungswasserwirtschaft," unpublished, KfW Entwicklungsbank, Frankfurt a.M.

KfW, 2006b."Ex Post Evaluierungskriterien und Ratingsystem für die deutsche bilaterale FZ. Eine Orientierung für die Ex Post Evaluierungen der KfW," unpublished, KfW Entwicklungsbank, Frankfurt a.M.

KfW, 2008. "Analyse des Zusammenhangs zwischen Projekterfolg und Risiko der FZ Projekte," unpublished, KfW Entwicklungsbank, Frankfurt a.M.

KfW, 2008a. "Interpretation Guide for Water Supply and Sanitation relating to the Water Sector Strategy," unpublished, KfW Entwicklungsbank, Frankfurt a.M.

KfW, 2009a. "LogFrame ländliche Wasserversorgung," unpublished, KfW Entwicklungsbank, Frankfurt a.M.

KfW, 2009b. "LogFrame städtische Wasserversorgung," unpublished, KfW Entwicklungsbank, Frankfurt a.M.

KfW, 2010. "Aus Fehlern lernen: Die Evaluierung von FZ-Maβnahmen," unpublished, KfW Entwicklungsbank, Frankfurt a.M.

KfW, 2011a. "Water Supply Umfrage – FZE," unpublished, KfW Entwicklungsbank, Frankfurt a.M.

KfW, 2011b. "Arbeitshilfe zur Planung und Durchführung von Ex-Post Evaluierungen durch KfW-eigene MitarbeiterInnen," unpublished, KfW Entwicklungsbank, Frankfurt a.M.

LONG, J.S., 1997. *Regression Models for Categorical and Limited Dependent Variables*, SAGE Publications, Thousand Oaks.

MAVROTAS, G., 2005. "Aid Heterogeneity: Looking At Aid Effectiveness from A Different Angle," *Journal of International Development*, 17(8), 1019–1036.

NDIKUMANA, L., 2012. "Applying evaluation to development and aid: Can evaluation bridge the micro-macro gaps in aid effectiveness?" Department of Economics and Political Economy Research Institute (PERI), University of Massachusetts, Amherst.

ROODMAN, D., 2007. "The Anarchy of Numbers: Aid, Development and Cross-Country Empirics," *The World Bank Economic Review*, 21(2), 255-277.

SVENSSON, J., 2006. "The institutional economics of foreign aid," *Swedish Economic Policy Review*, 13(2), 115-137.

WHO, 2003. *Domestic water quantity, service, level and health*, World Health Organization, Geneva.