How to assess the effectiveness of development aid projects: evaluation ratings versus project indicators

Author(s):
Metzger, Laura; Günther, Isabel

Publication Date:
2015-11

Permanent Link:

https://doi.org/10.3929/ethz-a-010549257 →

Originally published in:

Journal of International Development 27(8), http://doi.org/10.1002/jid.3189 →

Journal of International Development
J. Int. Dev. **27**, 1496–1520 (2015)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jid.3189

HOW TO ASSESS THE EFFECTIVENESS OF DEVELOPMENT AID PROJECTS: EVALUATION RATINGS VERSUS PROJECT INDICATORS

LAURA METZGER* and ISABEL GUENTHER
ETH Zurich, Zurich, Switzerland

Abstract: Most studies on project-based aid effectiveness rely on subjective evaluation ratings to measure projects' performance. Using the example of drinking water projects, this study compares evaluation ratings to objective, quantitative project indicators based on water supply to better understand the drivers of evaluation ratings. We find that evaluation ratings are only weakly correlated with improvements in water supply. Whereas the water supply-based project indicators are best explained by project design variables, evaluation ratings put more weight on project management and implementation. © 2015 UNU-WIDER. *Journal of International Development* published by John Wiley & Sons, Ltd.

Keywords: aid effectiveness; evaluation; performance indicators; water supply
JEL Classification: F35; O19

1 INTRODUCTION

Since the early years of development research, aid effectiveness has been hotly debated. Most research on aid effectiveness is undertaken at the macro-level (e.g. Burnside & Dollar, 2000 and Burnside & Dollar, 2004; Clemens *et al.,* 2012; Collier & Dehn, 2001; Collier & Dollar, 2002; Roodman, 2007) or the micro-level.[1] Macroeconomic studies, focusing on the effect of aid and country characteristics on growth, somewhat faded from the spotlight in the last 10 years. Microeconomic studies, and particularly randomized

*Correspondence to: Laura Metzger, ETH Zurich, Center for Development and Cooperation, Clausiusstrasse 37, 8092 Zurich, Switzerland.
E-mail: laura.metzger@nadel.ethz.ch

[1] The examples are numerous given that almost every aid intervention has to be evaluated separately. For an overview of randomized controlled trials, having become one of the most applied methods to study aid effectiveness at the micro level, for example www.povertyactionlab.org.

control trials, which study the social welfare impact of single aid interventions, gained in importance. The number of scientific microeconomic studies that are comparable across dependent and independent variables is still relatively small. Moreover, such studies do not allow for an analysis of the impact of country characteristics or project management variables. Cross-country studies on the effectiveness of (evaluated) aid projects can partly bridge this gap between macro-research and micro-research by simultaneously studying the influence of country-level and project-level factors on project performance (e.g. Kilby, 2000). However, cross-country research on the effectiveness of project-based aid is still limited and focuses on World Bank (WB) aid (e.g. Denizer *et al.,* 2013; Guillaumont & Laajaj, 2006; Isham & Kaufmann, 1999; Kilby, 2000).[2]

Most cross-country studies on project-based aid use ex-post evaluation ratings (Denizer *et al.,* 2013; Dollar & Levine, 2005; Guillaumont & Laajaj, 2006) or economic rates of return (e.g. Isham *et al.,* 1997; Isham & Kaufmann, 1999; Kilby, 2000) to measure project performance. Both indicators have drawbacks. Economic rates of return result from cost-benefit analyses and do not directly measure aid effectiveness.[3] Ex-post evaluation ratings are subjective assessments of the extent to which aid projects achieved their targeted development goals (Denizer *et al.,* 2013). Because evaluation ratings are subjective, their interpretability and comparability is limited. First, they are not standardized across donor organizations. Second, they may not even be comparable within an organization: evaluators follow common rules and guidelines, but are given flexibility in their rating decisions.[4] By aggregating projects from heterogeneous sectors, many studies further reduce the comparability and interpretability of the ratings. Moreover, because of the heterogeneity of different aid sectors, analyses on the drivers of project aid effectiveness can only include sector-unspecific project management variables.

Given these shortcomings, this study aims to better understand what evaluation ratings measure by comparing them to (more) objective project indicators. To achieve this goal, we concentrate on a single aid sector. Focusing on a single sector first allows us to define quantitative performance indicators. Second, it also allows us to use a larger set of project-related explanatory variables than previous studies: in addition to sector-unspecific *project management* variables, we use sector-specific *project design* variables, which are usually included in microeconomic Randomized Control Trial (RCT) studies. To our knowledge, this is the first study in the field to compare donor evaluation ratings with objective performance indicators and to use sector-specific project design characteristics as explanatory variables.

Our analysis is based on a representative sample of 150 drinking water projects that the German Development Bank (KfW), a large bilateral donor, financed on behalf of the German government. We chose water aid, first and importantly, because it offers clear quantitative project performance indicators. These indicators are applied to all KfW water projects (KfW, 2009a; KfW, 2009b) and could, in principle, be used for other donors' drinking water projects. Moreover, water aid was, and still is, a main activity of the German Development Cooperation.[5]

In addition to KfW's evaluation ratings, we use two indicators that are crucial for measuring the performance of drinking water projects: (i) *the relative change in water consumption in litres per capita and day* and (ii) *the relative change in the number of individuals with access to an improved water point* between project appraisal (i.e. baseline)

[2] The WB has taken a lead role in providing data from its operative business for research. In 2011, its Independent Evaluation Group published a data set with evaluation ratings for all project and programme assessments that had been carried out since the 1970s.

[3] By aid effectiveness, we mean the extent to which an individual project achieved its development goals.

[4] See the IEG evaluator manual: https://ieg.worldbankgroup.org/Data/reports/icr_manual.pdf.

[5] https://www.kfw-entwicklungsbank.de/Internationale-Finanzierung/KfW-Entwicklungsbank/Themen/Wasser/

and evaluation. In accordance with KfW's results chain,[6] these water supply indicators capture the extent to which the immediate project goal—to provide a defined target population with a sufficient amount of safe drinking water—was achieved at the time of evaluation (KfW, 2009a; KfW, 2009b). Improved drinking water supply is a necessary precondition for the overall welfare objective KfW targets: better health through reducing water-borne diseases. Given their relationship to the immediate and overall project goals, the supply indicators should be significantly positively correlated especially with the effectiveness evaluation rating, which measures the immediate project goal. The water supply indicators should further be important for measuring the performance of drinking water projects, because they conform to internationally agreed indicators. Target 7C of the Millennium Development Goals was measured by the 'proportion of population using an improved drinking water source'. The United Nation's *Right to Water* declaration demands that a nation's entire population has access to safe water and that each person is guaranteed a consumption of 50 to 100 L a day.[7]

Our main findings are the following. First, contrary to our hypothesis, we find a weak correlation between (overall and effectiveness) evaluation ratings and water supply indicators. Second, our results suggest that country characteristics and *project management* variables explain more of the variation in evaluation ratings, while *project design* variables explain more of the variation in water supply indicators. Third, we find that the same explanatory variables affect effectiveness ratings and water supply indicators differently, although they should conceptually be similar if not the same. This leads to different conclusions about what may drive increased aid effectiveness.

The rest of the paper is structured as follows. In Section 2, we review the relevant literature on project-based aid effectiveness. In Section 3, we present the data and the analysis approach. In Section 4, we present and discuss the results. Section 5 concludes. Robustness tests for our findings are provided in the Appendix.

2 LITERATURE REVIEW

Cross-sectional studies on project-based aid effectiveness can be roughly divided into two groups. Studies in the first group rely on country characteristics (Dollar & Levine, 2005; Dreher *et al.,* 2013; Guillaumont & Laajaj, 2006; Isham *et al.,* 1997; Isham & Kaufmann, 1999). Studies in the second group add sector-unspecific project management variables to the set of country characteristics to explain project performance (Denizer *et al.,* 2013; Dollar & Svensson, 2000[8]; Kilby, 2000). Except for Hemmer and Lorenz (2003) and Ivanova *et al.* (2001), all studies focus on WB aid and measure project performance with ERR or evaluation ratings.

Studies in the first group can be briefly summarized as follows. According to Dollar and Levine (2005) and Isham *et al.* (1997), good institutions (indicated by the rule of law, a pluralistic and free media, and civil and political rights) significantly raise the probability that a project is rated 'successful'. A complementary study by Dreher *et al.* (2013) examines if 'politically motivated aid', that is, aid that is allocated for political reasons and not partner country need, is less effective. They find that evaluation ratings are only

[6] The results chain ('logical framework') specifies how a project's inputs and activities lead to desired key objectives. Each key objective has a defined set of indicators.

[7] *Right to Water*: *Human Rights Fact Sheet No.35*, United Nations (2010).

[8] Dollar and Svensson (2000) analyse 220 World Bank structural adjustment programmes.

negatively affected by politically motivated aid if the partner country is economically vulnerable. Otherwise, politically motivated aid does not significantly affect the evaluation ratings. According to Guillaumont and Laajaj (2006), economic instability, approximated by the variability of a country's exports, significantly reduces the probability that a project is rated 'successful'. Moreover, they find that an increasing level of overall aid received by the partner country is negatively correlated with project success. The Hemmer and Lorenz (2003) study also underlines that economic stability affects evaluation ratings: a better macroeconomic environment, indicated by less inflation, a balanced budget and trade openness, is significantly positively correlated with project ratings.

Studies in the second group simultaneously consider country-level characteristics and sector-unspecific project management variables. Dollar and Svensson (2000) analyse the impact of programme preparation and supervision time and the number of conditions attached to a programme loan on project performance. They find that programme management variables explain little of the variation in the WB ratings. The quality of the local political economy, by contrast, has a significant impact on the success of structural adjustment programmes. Dollar and Svensson conclude that the WB has little influence on programme performance through monitoring and supervision if the partner country's environment is politically and economically risky. Similarly, Ivanova *et al.* (2001) find that the International Monetary Fund (IMF's) steering and management efforts have no significant effect on country programme outcomes. Political instability, political cohesion and the influence of political interest groups have a higher impact on programme success. Kilby (2000) investigates how supervision time affects the success of WB projects. In contrast to Dollar and Svensson (2000) and Ivanova *et al.* (2001), who focus on macro-economic programmes, Kilby (2000) finds that additional supervision time, when invested early in the project management process, is significantly correlated with a better rating of projects.

Denizer *et al.* (2013) consider a larger set of project management variables than previous studies. They find that early warning indicators issued by task managers to mark problematic projects are significantly negatively correlated with evaluation ratings. Project preparation and supervision costs are also significantly negatively correlated with the ratings, which contradict Kilby's (2000) results on supervision costs. Furthermore, Denizer *et al.* (2013) find that larger projects and projects with a longer implementation period get lower ratings. Last, the quality of the task manager is significantly positively correlated with project ratings. Last, they find that evaluation ratings are strongly positively correlated with Country Policy and Institutional Assessment scores. However, overall, they find that the explanatory power of country characteristics is fairly limited.

Many of the aforementioned studies explain a rather small share of the variation in ERR or evaluation ratings. Isham *et al.* (1997), for example, attain unadjusted R^2 values between 0.02 and 0.18, the adjusted R^2 of the linear probability models of Dollar and Svensson (2000) lies between 0.15 and 0.28 and the unadjusted R^2 of the main regressions of Denizer *et al.* (2013) ranges between 13 and 16 per cent.

3 DATA AND EMPIRICAL APPROACH

3.1 Data set

KfW's independent evaluation department (FZE hereafter) carries out all project evaluations. Before, FZE evaluated every completed project. Since 2007, it has evaluated

50 per cent of all projects. KfW uses a (by sector) stratified random sampling approach to draw a representative sample of the projects that are to be evaluated. The sampling procedure is monitored by an assessor who is external to FZE to ensure that no project is purposefully selected or excluded. Ten to fifteen per cent of yearly evaluations are commissioned to external consultants. The rest are carried out by KfW employees who were never involved in the management of the project they evaluate. The final evaluation report provides a qualitative assessment of project performance, supplemented with descriptive, quantitative data.

Our data set spans the years between 1965 and 2011. It is composed of 150 KfW drinking water projects, from 62 countries. The countries are listed in Table A1 in the Appendix. Out of the 272 drinking water projects that FZE evaluated between 1989 and 2011, we used those for which documentation was available.[9] The evaluation ratings (Section 3.2), water supply indicators (Section 3.3) and project-level covariates (Section 3.4) used in the analysis were manually extracted from a total of 300 project appraisal and evaluation reports.[10] A single report is, on average, about 50 pages. Project appraisal and evaluation reports are usually based on a one-week to two-week field trip to the project site. Desk reports are an exception.

3.2 Evaluation ratings

Prior to 2000, FZE assigned only overall project ratings. Between 2000 and 2006, it evaluated all projects with respect to their *significance* (equivalent to the OECD-DAC relevance criterion), *efficiency*, *effectiveness* and *impact*. Because the alignment of the evaluation criteria aligned with the OECD-DAC guidelines in 2006, projects have been evaluated with respect to five evaluation criteria: *relevance*, *efficiency*, *effectiveness*, *impact* and *sustainability*. *Sustainability* did not exist as a stand-alone criterion before 2006, but was treated as a cross-cutting issue.

The immediate goal of drinking water projects is to supply a defined target population with a sufficient amount of safe[11] drinking water, which is to be achieved, mainly, by investing in supply infrastructure. The extent to which the goal has been achieved is reflected in the *effectiveness* rating. The achievement of the immediate project goal is a necessary precondition for achieving the overall welfare objective captured by the *impact* rating: to improve the target population's health by reducing water-borne diseases. Ideally, improved drinking water supply and associated health benefits are sustained in the long-term. The *sustainability* rating indicates if a project meets this requirement. Furthermore, it is considered important that water suppliers cover the full costs of water provision through tariff revenues.[12] Full cost coverage, so the argument, guarantees the target group's long-term supply with improved drinking water (KfW, 2008). The

[9] We did not recover all reports. In the 1990s, KfW introduced a computer-based project management system, and a considerable number of reports were not digitalized. We investigated with the IT department if this happened for a particular reason, that is, whether we face a selection bias. This is apparently not the case.

[10] All information drawn from reports was explicitly stated in figures (including evaluation ratings). Subjective text interpretation was not necessary.

[11] 'Safe' means complying with World Health Organization guidelines.

[12] Full costs comprise system operation and maintenance costs and (re)-investments cost in physical infrastructure (KfW, 2008).

DOI: 10.1002/jid

efficiency[13] rating reflects if cost coverage is achieved. The *relevance* criterion evaluates whether the project addressed the development priorities of the target group, the partner country and the German government.[14]

Relevance, *impact*, *effectiveness* and *efficiency* are each rated on a scale of 1 (*very successful*) to 6 (*project failed*). *Sustainability* is rated on a scale of 1 (*very good sustainability*) to 4 (*insufficient sustainability*). Evaluators are supposed to rate the five sub-ratings separately and then combine them into an *overall evaluation rating* that ranges from 1 (*very successful*) to 6 (*project failed*). Similar to the WB's ratings, there are no predefined weights for the aggregation of the single criteria. To simplify the interpretation of the regressions coefficients (Section 4), we reversed the rating scale, such that 1 refers to *project failed* and 6 refers to *very successful*. The sub-ratings and the overall rating are explicitly stated in each evaluation report, in words and numbers. Subjective interpretations of the reports were not necessary.

Following previous studies, we measure project performance with the *overall* evaluation rating. In addition, we consider the *effectiveness* (sub)-rating, which measures the extent to which the immediate project goal has been achieved. KfW's results chain relates the *effectiveness* rating directly to the water supply indicators that we use for our analysis (Section 3.3): the *change in water consumption per capita in litres per day* and *the change in number of project beneficiaries with access to an improved water source* from project appraisal to project evaluation. Thus, the effectiveness rating and the water supply indicators measure the same concept. Given this close relationship, we focus our analysis on comparing the *effectiveness* rating with the water supply indicators. Data on the target group's health before and after an intervention are not available, so we do not separately analyse the *impact* and *sustainability* ratings.

3.3 Water supply indicators

The water supply indicators are constructed on the basis of quantitative data taken from appraisal and evaluation reports. Appraisal reports are prepared shortly before a project starts and contain two types of values: actual water supply levels at baseline (T_0) and targeted water supply levels to be achieved at the end of the project. Evaluation reports state the actual water supply levels at the time of evaluation (T_1). With this information at hand, we construct three water supply indicator categories as outlined in Table 1.

In a first step, we investigate if evaluators distinguish between these three conceptually different categories by testing which category is most strongly correlated with the evaluation ratings, especially the *effectiveness* rating (Section 4 and Table 2). However, in our main analysis (Section 4 and Tables 3–5), which compares what respectively drives evaluation ratings and water supply indicators, we focus on the *change in water consumption* and *change in population with access to an improved water source*. In our view, these quantitative indicators best reflect the performance of drinking water projects. The relative change in water supply between project start and end focuses on the benefits that the project generated for the

[13] Cost coverage is one of several components entering the efficiency criterion:https://www.kfw-entwicklungsbank.de/Internationale-Finanzierung/KfW-Entwicklungsbank/Evaluierung/Ex-Post-Evaluierungen/Schlüsselkriterien/

[14] For the interpretation of our results, it is important to note that none of the evaluation ratings is supposed to measure the quality of the organization's staff.

Table 1. Definitions of drinking water supply indicators

Performance indicators	Definition
(1) Supply levels at evaluation	
Water consumption lpcd in T_1	Water consumption in lpcd from an improved water source at evaluation (natural logarithm)
Population with access in T_1	Number of individuals with access to an improved water source at evaluation (natural logarithm)
(2) Relative change in supply levels between baseline and evaluation	
Change in water consumption lpcd	Change in daily water consumption (lpcd) from an improved water source (percentage change)
Change in population with access	Change in number of individuals with access to an improved water source (percentage change)
(3) Goal achievement at evaluation	
Goal achievement water lpcd in T_1	Actual supply (lpcd) divided by targeted daily water consumption (lpcd) from an improved water source at evaluation (percentage)
Goal achievement population with access in T_1	Actual divided by targeted number of individuals with access to an improved water source (percentage)

lpcd, litres per capita per day.

target group. In contrast, the status quo of water supply at evaluation does not assess improvements and might be highly correlated with baseline values. Goal achievement indicates whether planning values have been met. Because of long-time planning horizons, planning values are often obsolete when the project is evaluated. Goal achievement hence measures the quality of project planning rather than project performance.

Whereas KfW's project reports provide comprehensive project information, uncertainties about the precision of the water supply indicators remain. In case of limited data, project planners and evaluators sometimes approximate the values of the supply indicators (daily water consumption per capita in litres; number of target group individuals accessing an improved water point). However, planning and evaluation teams usually include engineers with long-standing (field) experience, trained in assessing the capacity of an installed water supply system with regard to the water quantity it produces and the population it serves. Hence, even if values might be noisy for certain projects, approximated water supply values should be within a reasonable range, and not be systematically biassed.

Data on chemical and/or microbiological drinking water quality are not available from the project reports. Hence, we cannot add water quality as an objective performance indicator (to be compared with the effectiveness ratings). We can only assess whether individuals have access to safe drinking water, not whether they consume chemically and microbiologically safe water. However, it is important to mention that at project finalization, the water quality of the source is tested and must be evaluated as safe. It is further important to note that the population-based supply indicator measures the number of targeted people that actually use a safe water source. Hence, quantitative supply indicators are more than a mere assessment of the physical infrastructure's functionality.

3.4 Explanatory variables

Country-level characteristics are represented by the following set of variables: GDP per capita indicates a country's development level; inflation indicates its macroeconomic

Table 2. Correlation between water supply indicators and evaluation ratings

	DV: effectiveness rating				DV: overall rating			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Total water supply from safe technology in T_1								
Water consumption ln(lpcd) in T_1	1.398^{*} (0.097)			0.652 (0.657)	1.157^{*}(0.083)			1.515 (0.179)
Water consumption $\ln(\text{lpcd})^2$ in T_1	-0.195^{*} (0.088)			−0.075 (0.686)	-0.170^{*} (0.053)			−0.195 (0.158)
Population (ln) with access in T_1	1.506 (0.147)			−0.576 (0.614)	1.330^{**}(0.026)			1.433^{**}(0.040)
Population (ln) with access^2 in T_1	−0.051 (0.242)			0.032	-0.046^{*}(0.063)			-0.049^{*}(0.085)
Change in water supply from safe technology from T_0 to T_1								
Change in water consumption lpcd		−0.189 (0.558)		0.043 (0.902)		−0.033 (0.891)		0.171(0.466)
Change in population with access		1.113^{***}(0.001)		1.131^{***}(0.003)		0.679^{**}(0.031)		0.586^{*}(0.075)
Achievement of target water supply from safe water technology in T_1								
Goal achievement water lpcd in T_1			−0.100 (0.639)				−0.097 (0.547)	
Goal achievement population with access in T_1			0.200^{*}(0.096)				0.222^{**} (0.035)	
Constant	−8.785 (0.169)	3.746^{***}(0.000)	4.016^{***}(0.000)	4.586 (0.549)	-6.904^{*} (0.063)	3.636^{***} (0.000)	3.857^{***} (0.000)	-9.122^{*} (0.052)
Observations	88	54	85	54	135	78	130	78
R^2	0.176	0.174	0.036	0.249	0.152	0.058	0.038	0.222
Adjusted R^2	0.136	0.141	0.013	0.153	0.126	0.033	0.023	0.156

Standard errors are in parentheses.
***Significance at 1 per cent level;
**significance at 5 per cent level;
*significance at 10 per cent level.
See Table 1 for an overview of all project performance indicators and their definitions. lpcd, litres per capita per day.

Table 3. Correlation between country characteristics and project performance

	DVs: change in water supply		DVs: evaluation ratings	
	Change in water consumption lpcd	Change in population with access	Effectiveness rating	Overall rating
	(1)	(2)	(3)	(4)
Average GDP per capita[a]	0.166* (0.056)	−0.107* (0.085)	−0.158 (0.406)	−0.182 (0.192)
Average bilateral aid per capita	0.003 (0.868)	0.000 (0.974)	−0.067** (0.035)	−0.063** (0.016)
Average inflation rate	−0.000 (0.783)	0.000 (0.551)	−0.001 (0.473)	−0.001** (0.030)
Conflict	−0.047 (0.581)	0.023 (0.705)	−0.299* (0.084)	−0.229* (0.089)
Polity index	−0.018* (0.063)	−0.002 (0.773)	0.045** (0.029)	0.028* (0.065)
Africa region	0.166 (0.299)	0.090 (0.416)	−0.062 (0.855)	−0.100 (0.692)
MENA region	0.090 (0.635)	0.136 (0.369)	0.938** (0.029)	0.423 (0.196)
Constant	−1.031* (0.097)	1.126** (0.012)	5.552*** (0.000)	5.608*** (0.000)
Observations	87	83	82	133
R^2	0.134	0.085	0.155	0.115
Adjusted R^2	0.057	−0.000	0.075	0.066

Standard errors are in parentheses.
***Significance at 1 per cent level;
**significance at 5 per cent level;
*significance at 10 per cent level.
[a]logarithm.
See Table 1 for an overview of all project performance indicators and their definitions. lpcd, litres per capita per day.

Table 4. Correlation between project management characteristics and project performance

	DVs: change in water supply		DVs: evaluation ratings	
	Change in water consumption lpcd	Change in population with access	Effectiveness rating	Overall rating
	(1)	(2)	(3)	(4)
Months past project closure	−0.002 (0.202)	−0.001 (0.422)	0.002 (0.706)	0.003 (0.306)
Project duration	0.002 (0.371)	0.002 (0.136)	−0.004 (0.369)	−0.008*** (0.008)
Second phase[a]	−0.045 (0.745)	−0.025 (0.825)	0.128 (0.673)	0.000 (0.999)
Third phase	0.749** (0.020)	0.442 (0.240)	−0.334 (0.659)	0.530 (0.441)
Risk evaluation in T_0	0.006 (0.948)	−0.021 (0.781)	−0.544** (0.012)	−0.514*** (0.000)
Total project costs in USD[b]	−0.029 (0.652)	−0.022 (0.677)	0.439*** (0.005)	0.233*** (0.024)
Partner country share in total project costs	−0.259 (0.329)	−0.018 (0.932)	−0.714 (0.316)	−0.312 (0.517)
Constant	0.744 (0.479)	0.868 (0.304)	−1.702 (0.489)	1.386 (0.258)
Observations	77	80	87	125
R^2	0.137	0.073	0.166	0.161
Adjusted R^2	0.050	−0.017	0.092	0.110

Standard errors are in parentheses.
***Significance at 1 per cent level;
**significance at 5 per cent level;
*significance at 10 per cent level.
[a]Phase 1 is the left out category;
[b]logarithm.
See Table 1 for an overview of all project performance indicators and their definitions. lpcd, litres per capita per day.

Table 5. Correlation between project design and project performance

	DVs: change in water supply		DVs: evaluation ratings	
	Change in water consumption lpcd	Change in population with access	Effectiveness rating	Overall rating
	(1)	(2)	(3)	(4)
Tariff per m^3 Consumed [a]	−0.096 (0.668)	−0.417** (0.020)	0.995** (0.039)	0.730** (0.032)
Mixed system[b]	0.008 (0.987)	0.127 (0.651)	−1.108 (0.123)	−0.754 (0.114)
Piped system	0.640*** (0.001)	0.184 (0.268)	−1.003** (0.047)	−0.280 (0.426)
Urban	−0.009 (0.918)	−0.083 (0.169)	0.108 (0.621)	0.143 (0.343)
User committee	0.536*** (0.002)	0.057 (0.692)	0.279 (0.544)	0.806*** (0.006)
Wastewater component	−0.122 (0.199)	−0.018 (0.813)	0.205 (0.419)	0.024 (0.896)
Hygiene promotion	0.286*** (0.004)	0.239*** (0.003)	0.021 (0.945)	−0.083 (0.688)
Institutional support	0.010 (0.917)	−0.129* (0.094)	0.351 (0.269)	0.054 (0.793)
Low water resources	0.100 (0.285)	−0.089 (0.229)	−0.603** (0.021)	−0.330* (0.067)
Constant	−0.536*** (0.006)	0.567*** (0.001)	4.491*** (0.000)	3.781*** (0.000)
Observations	89	89	82	134
R^2	0.310	0.309	0.134	0.186
Adjusted R^2	0.232	0.230	0.026	0.085

Standard errors are in parentheses.
***Significance at 1 per cent level;
**significance at 5 per cent level;
*significance at 10 per cent level.
[a] logarithm;
[b] non-piped system is the left out category.
See Table 1 for an overview of all project performance indicators and their definitions. lpcd, litres per capita per day.

stability. We also include the absolute and squared value of bilateral aid per capita to test if total aid influences project outcomes. The cost of managing high aid volumes might be large and negatively affect the performance of single projects. For example, if administrative capacity is limited, an increasing number of projects might entail less time dedicated to the management, implementation and monitoring of each project. The variables are taken from the World Development Indicators (2013). We computed the yearly average of GDP per capita, bilateral aid per capita and inflation rate over the length of the project, that is, from its start to its evaluation.

We also included a categorical variable from the UCDP/PRIO data set that indicates if an armed conflict took place between project start and evaluation. It is set to 1 if there was a minor conflict (25 to 999 battle-related deaths), set to 2 if there was a war (over 999 battle-related deaths) and 0 if there was no conflict.[15] The influence of a country's regime type on project performance is captured by the polity2 indicator, published by the Centre for Systemic Peace.[16] The indicator ranges from −9 to 10. Higher scores represent a more democratic regime. We use the polity2 value at project start. We further consider region-fixed effects for sub-Saharan Africa (poorest region) and the Middle East (region with highest water scarcity). We prefer region effects to country-fixed effects to save degrees of freedom. Our set of explanatory variables is an extension of the country characteristics used by Denizer *et al.* (2013).

[15] http://www.pcr.uu.se/research/ucdp/datasets/ucdp_prio_armed_conflict_dataset/
[16] http://www.systemicpeace.org/polityproject.html

Project-level characteristics were extracted from project appraisal and evaluation reports. First, we compiled a list of what we call *project management* variables. As in previous studies, these variables are sector unspecific. Project size is indicated by total project costs (constant dollars). We also account for the partner country's share in total project costs, which may indicate the financial capacity of the partner country, and/or the policy relevance of drinking water issues. Furthermore, we consider project duration and the time in months that elapsed between project closure and evaluation. We also consider project phases. A project phase corresponds to a full-fledged project and is separately evaluated. Finally, the project planner's perceived risk of project failure (before project start) is captured by an ordinal variable ranging from 1 (*low risk*) to 4 (*very high risk*).

Besides project management variables, we consider *project design* variables that are specific to drinking water projects. To our knowledge, we are the first to simultaneously use *project design* variables, *project management* variables and country characteristics in a cross-sectional study on project-based aid effectiveness. We consider if the project was located in a rural or urban area and if a water user committee was put in place to operate and maintain the water supply system (this is often performed in rural project areas underserved with public services). Furthermore, we consider if support activities were undertaken to institutionally strengthen the water supplier. We also take into account if a hygiene campaign was part of the project activities. Hygiene campaigns are mainly integrated into rural projects, because the rural populations' knowledge about health-related advantages of safe water consumption is often insufficient.[17] We also consider which type of supply system is used to serve the target population: non-piped (hand pumps), mixed (piped and non-piped) or piped system.

Another important variable is the local tariff that is charged per metre cubed of consumed water.[18] Given the relationship between tariffs and cost-coverage described earlier, and the alleged influence of cost coverage on project sustainability, this variable could positively affect the *overall* rating. However, higher tariffs can negatively affect consumption levels and access rates and therefore negatively affect the supply indicators and the *effectiveness* rating. Moreover, we add a dummy to indicate if wastewater treatment was part of the project. Wastewater treatment does not directly affect drinking water supply, but may have an indirect positive effect on it, especially in urban areas producing large wastewater quantities: treated wastewater affects the groundwater volume available for drinking water purification.[19] Last, a dummy accounts for a country's internal water resources. The data are taken from the FAO AQUASTAT databank.[20]

It should be noted that, although we use a more comprehensive set of project-level variables than previous comparable studies, the information is still too limited to concentrate the analysis on the question how project design and management drive the performance of drinking water projects. However, as indicated earlier, our research objective is 'only' to better understand what donor evaluation ratings imply and measure.

[17] Hygiene campaigns aim to sensitize the target population for choosing safe technologies (e.g. hand pumps) over unsafe sources (e.g. unprotected surface water). Moreover, they intend to instruct households how to transport, store and handle drinking water to prevent contamination with bacteria.

[18] We have information on the tariff at two points in time: appraisal and evaluation, but more missing values at the time of appraisal. To not lose more observations for our already small sample, we use the tariff at evaluation.

[19] http://www.un-igrac.org/dynamics/modules/SFIL0100/view.php?fil_Id = 173

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

3.5 Estimation approach

Our analysis consists of two parts. In the first part (Table 2), we analyse the relationship between the ratings and the supply indicators by separately regressing the *effectiveness* and *overall rating* on the water supply indicators defined in Table 1. Because the water supply indicators capture to what degree the immediate project goal was achieved, they should be significantly and positively correlated with both ratings. However, for the reasons discussed in Section 3.2, we expect that the supply indicators are more strongly correlated with the *effectiveness rating* and moreover hypothesize that they explain a very large share of its variation. We also test which water supply indicator category is weighted the most when evaluators assign a rating: the total supply levels at evaluation, the relative change in supply between baseline and evaluation or the goal achievement at evaluation.

In the second and main part of our analysis, we try to better understand what respectively drives evaluation ratings and water supply indicators. For the previously discussed reasons, we focus on comparing the *effectiveness* rating with the supply indicators. We separately regress the supply indicators and the ratings on our set of country-level and project-level characteristics (Table 3–5). Our hypothesis is that the explanatory variables affect the *effectiveness* rating and the water supply indicators in a very similar fashion. There should be fewer similarities between the water supply indicators and the *overall* rating, because the latter captures factors beyond water supply. Moreover, because they measure different concepts, we should observe differences between the *overall* and the *effectiveness* rating. Furthermore, we are interested in which set of explanatory variables best explains the variation in our dependent variables. This is an empirical question.

We work with a pooled cross-section of 150 water projects. Because our sample is composed of a randomly drawn sample of finalized projects, we can reasonably assume that our cross section is independently pooled. Because of missing data, especially for the baseline values of the water supply indicators, the sample size varies with the model specifications. It is mostly smaller than our full sample of 150 projects. We verified that the data is missing at random. Test statistics are available from the authors upon request. Nevertheless, the small sample size decreases the precision of the point estimates as standard errors become larger and confidence intervals widen. However, despite the small sample size our results are fairly robust to different specifications. The robustness tests are separately shown in the Appendix (Table A5 and A6). It should be noted that related studies are based on similar sample sizes: Dollar and Levine (2005) have a sample of 52 to 90 observations; Ivanova *et al.* (2001) have a sample size of 55 to 61.

As with previous studies in the field, our cross-sectional study is based on non-randomized observational data that only allows for a before–after comparison of project indicators. Causal inference is only possible to a very limited extent. However, as we focus on comparing evaluation ratings with project indicators, endogeneity issues should be less severe than in studies seeking to identify what determines aid effectiveness. All regressions are estimated with ordinary least squares (OLS).[21]

[21] The results for the evaluation ratings do not change in sign and significance if estimated with an ordered logit. Hence, for easier comparison across regressions, we present the OLS estimation.

4 RESULTS

Table 2 shows the correlation between the water supply indicators and both evaluation ratings. Columns (1) to (4) refer to the *effectiveness* rating. Columns (5) to (8) refer to the *overall* rating. The column pairs (1)/ (5), (2)/ (6) and (3)/ (7) respectively correspond to the (i) total water supply level at evaluation, (ii) change in water supply between appraisal (baseline) and evaluation and (iii) goal achievement at evaluation. The water consumption and population-based supply indicators appear in the same regression, because evaluators consider them simultaneously in the evaluation process. High collinearity between the explanatory variables is not an issue (Table A4). In columns (4) and (8), we combine the supply levels and the change in supply.

The squared terms of *daily per capita safe water consumption* and *number of people with access to an improved water source* in the regressions in columns (1), (4), (5) and (8) account for non-linear effects. KfW defines 'adequate' consumption levels for the target group, mainly depending on the project area characteristics and the supply system that is installed or rehabilitated. These consumption levels are based on a pre-defined scale, ranging from 20 L (rural area/non-piped system) to 180 L (urban area/piped system) per capita and day. Exceeding consumption targets can, in certain cases, negatively affect the ratings. For example, if 'excessive' water consumption results from systemic water losses, a project's effectiveness may be compromised. We indeed find non-linear effects: higher water consumption levels lead to a better rating, but only up to a certain point. A similar statement applies to the number of people in the target group accessing an improved water source.

Our results further indicate that the population-based indicators are more strongly correlated with the evaluation ratings, while water consumption indicators only weakly influence them. Consumption levels at evaluation are in some specifications significantly correlated with the ratings; relative changes in water consumption never have a significant influence on the ratings. Based on the results in Table 2, we cannot say which supply indicator category has the strongest impact on the ratings. It seems that supply at evaluation and change in supply between project appraisal and evaluation are about equally important to evaluators, whereas goal achievement is less important. Moreover, we observe little difference between the *effectiveness* and the *overall* rating. The supply indicators affect both ratings in a similar way: the values of the coefficients are very close to each other.[22] Moreover, the adjusted R^2 values of the regressions based on *effectiveness* are low given that this rating should capture improvements in safe water supply. Hence, contrary to our hypothesis, water supply indicators are weakly correlated with, and explain little of the variation in the *effectiveness* rating. Apparently, evaluators do not attach high weight to information about water supply in assessing the achievement of the immediate project goal.

Next, we address the second and main part of our analysis. As stated earlier, we focus on *the change in water consumption* and the *change in population with access to an improved source* as dependent variables, because change scores reflect the impact of drinking water projects better than supply levels at evaluations and goal achievement. We separately regress the ratings and supply indicators on our three sets of explanatory variables: (a)

[22] This holds when we restrict the sample in column (8) to the 54 observations in column (4). Results are available from the authors upon request.

DOI: 10.1002/jid

country characteristics (Table 3), (b) *project management* characteristics (Table 4) and (c) *project design* characteristics (Table 5).

Table 3 shows that the occurrence of a conflict negatively affects both evaluation ratings, but has no significant influence the change in water supply over time. Similarly, a better political environment positively influences both ratings, whereas it hardly affects water supply (the negative correlation with the *change in per capita consumption levels* (column 1) disappears, once we add project-level variables to the regression (Table (A5)). Furthermore, we observe that increasing levels of bilateral aid per capita are significantly negatively correlated with both ratings.[23] This result parallels the finding of Guillaumont and Laajaj (2006) that increasing levels of Official Development Assistance, measured in per cent of GDP, reduce a project's success probability. However, this does not apply to the supply indicators. The results suggest that a country's political–institutional context influences the evaluation ratings but not the water supply indicators. On one hand, the political–institutional context may influence the overall project performance, which the more narrowly defined water supply indicators do not capture. That would explain the differences between the supply indicators and the *overall* rating, but not the differences between the supply indicators and the *effectiveness* rating. An alternative explanation might be that evaluators overestimate the influence of a country's political–institutional context, because this is often in the foreground of daily business operations.

Furthermore, we observe that a higher GDP per capita is positively correlated with the *change in per capita consumption levels* and negatively correlated with the *change in population with access to an improved s*ource. In (lower-income) middle-income countries, a relatively higher share of the population might already have access to safe water technologies, so substantial increases in the population served are more difficult to achieve. At the same time, water consumption levels may rise, even at higher income levels. As Table 3 shows, GDP per capita is not significantly correlated with the evaluation ratings. However, if we add the project-level variables to the regression, the correlation between GDP per capita and the *overall* rating is negative and significant (Table A5, column 4). It could be that we observe a negative relationship with the *overall* rating, because the impact of a drinking water project is perceived to be higher in poorer countries.

A last important observation is that the country-level characteristics explain a very small share of the variation in project performance, as the low adjusted R^2 values indicate. This result might be specific to the drinking water sector, which is probably less directly affected by macroeconomic factors than other aid sectors such as micro-crediting. However, Guillaumont and Laajaj (2006) and Denizer *et al.* (2013), who pool heterogeneous sectors, also find that country characteristics explain a small share of the variation in the WB ratings.

The results in Table 4 show that projects with longer-than-planned implementation periods tend to get a lower *overall* rating. Denizer *et al.* (2013) also find that longer implementation periods of WB projects lead to lower ratings. However, we use the difference between planned and actual implementation period. Denizer *et al.* (2013) use the implementation period. Hence, in our case it seems that 'bad' planning is penalized by a lower rating. This interpretation is misleading if unobserved project quality jointly determines project prolongation and lower ratings. In that case, the point estimate of

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

project duration would be biassed downwards. However, when we use water supply indicators as a measure of project performance, we do not find evidence that unplanned project prolongation leads to lower performance.

The rating-based regressions support another key result found in earlier studies: a high perceived risk of project failure is significantly negatively correlated with the ratings. Evaluators seem to be more likely to give lower ratings if a project was initially classified as risky. The risk variable is insignificant when project performance is measured with the water supply indicators. Another finding (partly) in line with previous work is that project costs are positively correlated with evaluation ratings (Denizer *et al.* (2013) find the opposite). Kilby (2000) suggests that this might be grounded in the fact that loans in ‘well-established’ sub-sectors are larger or because committed borrowing governments receive larger loans. We do not find a significant relationship between loan size and supply indicators.[24] The third phase of a project is the only *project management* variable that is significantly correlated with improvements in water supply. Possibly, later phases are more successful because of learning effects, or because only successful projects enter a subsequent phase.

Comparing columns 3 and 4 of Table 4 shows that the *effectiveness* and the *overall* rating are very similarly affected by the explanatory variables, although the two ratings measure different dimensions of project performance. In contrast, there are few similarities between the *effectiveness* rating and the water supply indicators, even though both should measure improvements in water supply. Last, the low adjusted R^2 values indicate that *project management* variables explain very little the variation in water supply indicators.

Turning to *project design* variables, Table 5 shows that the local tariff per cubic metre of consumed water is significantly positively correlated with both evaluation ratings. That a higher tariff leads to a better *overall* rating fits well with the claim that cost coverage through tariffs is vital for sustainable drinking water provision. However, as one would expect, a higher tariff is significantly negatively correlated with the *change in the number of people accessing a safe water source*. The difference in results between the *overall* rating and the supply indicators seems reasonable: in contrast to the *overall* rating, supply indicators do not capture concerns about sustainability. This argument does not hold for the *effectiveness* rating, which is also positively affected by higher tariffs.[25]

Similarly, hygiene promotion has a significant positive impact on both water supply indicators. It seems that hygiene promotion successfully sensitizes the target groups to the benefits of safe drinking water consumption. The question is why this variable does not have a positive and significant effect on the evaluation ratings. Because there are presumably no concerns about sustainability, it is not clear why hygiene promotion should not positively affect the ratings. The reason might be that these ‘accompanying’ activities, as KfW calls them, are usually not considered during the evaluation process.

Furthermore, we find that working with a piped system reduces the effectiveness rating. Piped systems in developing countries are often characterized by systemic water losses. As mentioned earlier, these losses can negatively affect a project’s sustainability, and the effectiveness rating seems to capture that. In contrast, the having a piped system in place

[24] When we include the baseline values in the regression, we find mixed evidence on the relationship between project size and water supply indicators: project size is significantly positively correlated with the change in the number of people accessing a safe source, but again not significantly correlated with the change in consumption levels. See Table A6 in the Appendix.

[25] One report points out that increased tariffs likely excluded poor individuals from safe water usage. Yet, the report emphasizes the advantages of a higher tariff, namely, improved cost coverage.

Table 6. Variance explained by country, project management and project design

	Change in water consumption lpcd	Change in population with access	Effectiveness rating	Overall rating
	(1)	(2)	(3)	(4)
Country characteristics	−0.016	0.108	0.120	0.090
Project management	0.025	0.013	0.071	0.113
Project design	0.230	0.153	0.000	0.006

lpcd, litres per capita per day.

is significantly positively correlated with the *change in per capita safe water consumption*. This supports the finding from micro-level research that water consumption levels only increase significantly with piped water systems (e.g. Devoto *et al.,* 2013; WHO, 2003).

Finally, and in line with Table 3 and 4, the explanatory variables affect both evaluation ratings similarly, despite the fact that the overall rating should provide a more comprehensive evaluation of the project (including its efficiency, impact, sustainability and relevance) than the effectiveness rating. The adjusted R^2 values in column (3) and (4) show that *project design* variables explain very little of the variation in the evaluation ratings.

To conclude, our results suggest several things. First, in some cases, we find parallels between the *effectiveness* rating and the supply indicators. However, we also find important differences, as the coefficients for the tariff and for hygiene promotion illustrate. The positive relationship between tariffs and both evaluation ratings is likely grounded in KfW's idea of a good project, namely, that a water project is sustainable when water provision costs are covered through tariffs (not subsidies). However, the water supply indicators show that higher tariffs are associated with lower consumption levels, which may negatively affect the improvement of the supply situation and hence the health impact of a project. These differences are problematic when policy recommendations on how to improve the design of drinking water projects are to be given. Second, that we failed to observe noticeable differences between the *effectiveness* and the *overall* evaluation rating strongly suggests that they are co-determined. A possible explanation is that evaluators first decide on the overall rating and then assign the sub-ratings. This in turn suggests that the sub-ratings do not measure what they are supposed to measure.

Table 6 shows how much of the variance in each performance indicator is respectively explained by *country* characteristics, *project management* characteristics and *project design* characteristics. For each indicator, the comparisons across variable sets are based on the same sample size. Based on the adjusted R^2, we conclude that evaluation ratings are mostly affected by country characteristics (*effectiveness* rating) and *project management* (*overall* rating) variables, whereas both water supply indicators are best explained by *project design* variables.

5 DISCUSSION AND CONCLUSION

On the basis of 150 drinking water projects financed by the German government, we aimed to better understand the drivers of evaluation ratings. To answer this question, we focused on a single aid sector, which allowed us to compare subjective evaluation ratings with quantitative and (more) objective water supply indicators. In addition, the single-sector approach enabled us to use a richer set of project level variables than comparable previous studies.

First, contrary to our hypothesis, we find that the water supply indicators are only weakly correlated with the *effectiveness* rating. Second, country characteristics and *project management* variables best explain the variation in the evaluation ratings, while *project design* variables best explain the variation in the *change in water supply*. Country characteristics and *project management* variables are probably more important for the ratings, because evaluators put more weight on a country's institutional–political context, and the implementation process itself than on project design. Third, we find differences in how evaluation ratings and supply indicators are affected by the explanatory variables. As regards the *overall* rating, the differences can be (partly) explained with the fact that the *overall* rating measures factors beyond water supply. However, this explanation does not apply to the *effectiveness* rating, which should be very similar to the supply indicators. A possible explanation for this discrepancy is that the *effectiveness* rating is not autonomous, but (co)-determined by the *overall* rating. Hence, we would have to give different policy recommendations to improve the effectiveness of drinking water projects, depending on the performance measure we use. Our results suggest that, where possible, donors should use objective project performance indicators in addition to subjective ratings, to increase the transparency and accuracy of donor-led project evaluation and of cross-sectional studies on project aid. Ideally, the results of studies on project-based aid that use observational data could be compared with the results of micro-level studies (RCTs) that use experimental data. Such comparisons can facilitate the validation of research findings on the same topic across different methods and may help closing the 'micro–macro gap' in the aid effectiveness debate.

Last, it is important to emphasize the limitations of our study. Despite the fact that we use a larger set of project-level variables than previous studies in the field, the explanatory power of our models is low. Moreover, our data is too limited to give an in-depth answer to the question of *how* country characteristics, donor performance and project design drive the effectiveness of project-based aid. In order to be able to do such analyses (and improve studies like this one), donors need to collect, store and publish (more) data suited for a systematic quantitative analysis.

ACKNOWLEDGEMENTS

We thank the German Development Bank (KfW) for supporting our research. We especially thank Eva Terberger, Jan Schrader and Katja Michaelowa for highly valuable comments on earlier drafts of this paper. Any errors remain the sole responsibility of the authors.

REFERENCES

Burnside C, Dollar D. 2000. Aid, policies and growth. *American Economic Review* **90**(4): 847–868. DOI:10.1257/aer.90.4.847.

Burnside C, Dollar D. 2004. Aid, policies and growth: revisiting the evidence. Policy Research Working Paper 3251. The World Bank: Washington, D.C., DOI: 10.1596/1813-9450-3251

Clemens M, Radelet S, Bhavnani R. 2012. Counting chickens when they hatch: timing and the effects of aid on growth. *The Economic Journal* **122**(561): 590–617. DOI:10.1111/j.1468-0297.2011.02482.x.

Collier P, Dehn J. 2001. Aid, shocks, and growth. Policy Research Working Paper 2688. The World Bank: Washington, D.C., DOI: 10.1596/1813-9450-2688

Collier P, Dollar D. 2002. Aid allocation and poverty reduction. *European Economic Review* **46**: 1475–1500. DOI:10.1596/1813-9450-2041.

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* **105**: 288–302. DOI:10.1596/1813-9450-5646.

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. DOI:10.3386/w16933.

Dollar D, Levine V. 2005. Sowing and reaping: institutional quality and project outcomes in developing countries. Policy Research Working Paper 3524. The World Bank, Washington D. C. DOI: 10.1596/1813-9450-3524

Dollar D, Svensson J. 2000. What explains the success and failure of structural adjustment programs? *The Economic Journal* **110**(446): 894–917. DOI:10.1111/1468-0297.00569.

Dreher A, Klasen S, Vreeland J, Werker E. 2013. The costs of favoritism: is politically-driven aid less effective? *Economic Development and Cultural Change* **62**(1): 157–191. DOI:10.1086/671711.

Guillaumont P, Laajaj R. 2006. When instability increases the effectiveness of aid projects. Policy Research Working Paper 4034. The World Bank, Washington D.C. DOI: 10.1596/1813-9450-4034

Hemmer H, Lorenz A. 2003. What determines the success or failure of German bilateral financial aid? *Review of World Economics* **139**(3): 507–549. DOI:10.1007/BF02659673.

Isham J, Kaufmann D, Pritchett L. 1997. Civil liberties, democracy, and the performance of government projects. *World Bank Economic Review* **11**(2): 219–242. DOI:10.1093/wber/11.2.219.

Isham J, Kaufmann D. 1999. The forgotten rationale for policy reform: the productivity of investment projects. *Quarterly Journal of Economics* **114**(1): 149–184. DOI:10.1162/003355399555972.

Ivanova A, Mayer W, Mourmouras A, Anayiotos G. 2001. *What Determines the Implementation of IMF-Supported Programs?* IMF Working Paper. 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. DOI:10.1016/S0304-3878(00)00082-1.

KfW. 2008. Interpretation guide for water supply and sanitation relating to the water sector strategy. Unpublished internal document. KfW Entwicklungsbank, Frankfurt a.M.

KfW. 2009a. LogFrame ländliche Wasserversorgung. Unpublished internal document. KfW Entwicklungsbank, Frankfurt a.M.

KfW. 2009b. LogFrame städtische Wasserversorgung. Unpublished internal document. KfW Entwicklungsbank, Frankfurt a.M.

Roodman D. 2007. The anarchy of numbers: aid, development and cross-country empirics. *The World Bank Economic Review* **21**(2): 255–277. DOI:10.1093/wber/lhm004.

WHO. 2003. *Domestic Water Quantity, Service, Level and Health*. World Health Organization: Geneva.

APPENDIX

Table A1. Country coverage of KfW drinking water projects

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 litres				
Water Consumption lpcd in T_0	93	57.95	0.00	226.00[a]
Water Consumption lpcd in T_1	139	66.58	4.00	300.00
Change in Water Consumption lpcd	93	0.17	−0.98	1.00
Goal Achievement Water Consumption in T_1	137	0.94	0.20	3.60
Population with safe water access				
Population With Access in T_0	99	243’898	0.000	6’700’000
Population With Access in T_1	139	401’551	2’364	16’000’000
Change in Population With Access	96	0.499	−1.571	1.000
Goal Achievement Population With Access in T_1	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[b]
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
Partner Country 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 T_1 (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	0.34	0	1

[a] We excluded one project in Samoa, Turkey, Albania, and Egypt from the summary statistics for *water lpcd T_0_baseline*, *water lpcd T_1_actual*, *water lpcd T_1_target*. These projects are outliers since their goal was to reduce “excessive” water consumption.

[b] The summary statistic for this variable includes very high inflation values for Congo, Dem. Rep. The median is at 7.51.

Table A3. Correlation between Sub-ratings and overall evaluation rating

	DV: Overall Evaluation Rating					
	(1)	(2)	(3)	(4)	(5)	(6)
Effectiveness Rating	0.875*** (0.000)					0.454*** (0.000)
Efficiency Rating		0.736*** (0.000)				0.238*** (0.000)
Impact Rating			0.827*** (0.000)			0.384*** (0.000)
Sustainability Rating				0.812*** (0.000)		
Relevance Rating					0.827*** (0.000)	–0.041 (0.620)
Constant	0.431** (0.036)	1.230*** (0.000)	0.480* (0.073)	0.820 (0.170)	0.267 (0.438)	–0.214 (0.267)
Observations	93	93	89	35	91	86
R-squared	0.785	0.605	0.686	0.488	0.588	0.887
Adjusted R^2	0.782	0.600	0.683	0.473	0.584	0.881

Notes: standard errors are in parentheses. *** $^{(**)}$ $^{(*)}$ denotes significance at the 1 (5) (10) per cent level. The dependent variable is a rating scale ranging from 1 (“project failed”) to 6 (“very successful”).

Table A2. shows that each sub-rating (effectiveness, efficiency, impact, relevance and sustainability) is, as expected, significantly positively correlated with the overall rating. The better a project performed in any of the areas captured by the single ratings, the higher the overall rating. In the regression presented in column (6) we consider the sub-ratings simultaneously. Sustainability is excluded, because the number of observations would otherwise be reduced to 32. We observe that the size of the coefficients of the sub-ratings reduces, which shows that they are highly correlated. Relevance turned insignificant, which implies that the relevance sub-rating is not correlated with the overall ratings. KfW aligned its evaluation criteria with the OECD/DAC criteria in 2006. It is therefore interesting to examine how the relative sub-ratings’ influence on the overall rating changed after this date. Using a Chow test, we analyse if the regression coefficients in column (6) are the same before and after 2006, or if the revision of the criteria evoked a structural change in the evaluation system. We do not find evidence for a structural break: 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”. This result is plausible, since the alignment with the OECD/DAC criteria did not lead to a major revision of KfW’s evaluation criteria, at least in the drinking water sector. It rather extended them by giving sustainability the status of a stand-alone criterion. Moreover, the quantitative and qualitative indicators on which KfW is relying to assess the performance of drinking water aid projects, and which enter the ratings, did not change after 2006.

Table A4. Correlation matrix of project performance indicators

	Water Consumption lpcd in T_1	Population With Access in T_1	Change in Water Consumption lpcd	Change in Population With Access	Goal Achievement Water lpcd in T_1	Goal Achievement Population With Access in T_1
Water Consumption lpcd in T_1	1.00					
Population With Access in T_1	0.32 (0.00)	1.00				
Change in Water Consumption lpcd	−0.10 (0.40)	−0.06 (0.58)	1.00			
Change in Population With Access	−0.30 (0.01)	0.02 (0.87)	0.06 (0.61)	1.00		
Goal Achievement Water lpcd in T_1	0.41 (0.00)	−0.03 (0.79)	0.42 (0.00)	−0.19 (0.10)	1.00	
Goal Achievement Population With Access in T_1	0.09 (0.44)	0.24 (0.04)	−0.08 (0.51)	0.27 (0.02)	−0.09 (0.45)	1.00

Note.

**denotes significance at the 5 per cent level or less.

See Table 1 for an overview of the dependent variables and their definitions.

Table A5. Correlation between full Set of explanatory variables and project performance

	DV: Change in Water Supply		DV: Evaluation Ratings	
	Change in Water Consumption lpcd	Change in Population With Access	Effectiveness Rating	Overall Rating
	(1)	(2)	(3)	(4)
		Country Characteristics		
Average GDP Per Capita[a]	0.350*** (0.001)	−0.310*** (0.000)	−0.303 (0.289)	−0.600*** (0.002)
Average Bilateral Aid Per Capita	−0.005 (0.803)	0.038** (0.042)	−0.046 (0.344)	−0.020 (0.589)
Average Inflation Rate	−0.000 (0.963)	−0.000 (0.751)	−0.003 (0.118)	−0.001 (0.542)
Conflict	0.054 (0.564)	0.088 (0.213)	−0.212 (0.367)	−0.067 (0.669)
Polity Index	−0.017 (0.112)	0.003 (0.745)	0.051** (0.031)	0.033* (0.065)
Africa Region	0.133 (0.437)	−0.120 (0.352)	−0.483 (0.275)	−0.612* (0.054)
MENA Region	0.281 (0.243)	0.102 (0.620)	1.220** (0.032)	0.794* (0.064)
		Project Management Characteristics		
Months Past Project Closure	−0.002 (0.235)	−0.001 (0.460)	0.005 (0.352)	−0.002 (0.595)
Project Duration	0.001 (0.532)	0.004** (0.018)	−0.000 (0.964)	−0.006* (0.091)
Second Phase[b]	0.019 (0.897)	−0.177 (0.109)	−0.065 (0.869)	−0.012 (0.965)
Third Phase	−0.850* (0.053)	0.912** (0.034)	−0.446 (0.652)	0.772 (0.319)
Risk Evaluation in T_0	−0.062 (0.594)	−0.155* (0.081)	−0.536 (0.104)	−0.644*** (0.001)
Total Project Costs in USD[c]	−0.092 (0.192)	0.047 (0.350)	0.255 (0.195)	0.198 (0.105)
Partner Country Share in Total Project Costs	−0.529 (0.134)	0.517* (0.054)	0.189 (0.850)	0.433 (0.498)
		Project Design Characteristics		
Tariff Per m^3 Consumed [d]	−0.532* (0.097)	−0.204 (0.437)	0.859 (0.224)	0.989* (0.054)
Mixed System[e]	0.141 (0.776)	0.314 (0.322)	−1.017 (0.226)	−0.493 (0.357)
Piped System	0.965*** (0.002)	−0.036 (0.882)	−1.289* (0.051)	−0.829* (0.065)
Urban	0.079 (0.520)	−0.109 (0.170)	0.037 (0.887)	0.028 (0.879)
User Committee	1.049*** (0.000)	−0.299 (0.164)	m (0.743)	−0.014 (0.971)
Wastewater Component	0.070 (0.566)	−0.056 (0.524)	−0.030 (0.915)	−0.191 (0.360)
Hygiene Promotion	0.412*** (0.002)	0.201** (0.049)	0.177 (0.632)	−0.052 (0.830)
Institutional Support	0.125 (0.351)	−0.161* (0.081)	−0.137 (0.706)	−0.035 (0.883)
Low Water Resources	−0.000 (0.999)	−0.156 (0.131)	−0.766** (0.036)	−0.372 (0.146)
Constant	−1.586 (0.233)	2.132** (0.025)	4.367 (0.260)	7.241*** (0.002)
Observations	68	70	70	100
R−squared	0.539	0.580	0.469	0.402
Adjusted R^2	0.298	0.370	0.203	0.221

Notes: standard errors are in parentheses; *** (**) (*) denotes significance at the 1 (5) (10) per cent level;
[a]logarithm;
[b]phase 1 is the left out category;
[c]logarithm;
[d]logarithm;
[e]non-piped system is the left out category.
See Table 1 for an overview of all project performance indicators and their definitions.

published by John Wiley & Sons, Ltd. DOI: 10.1002/jid

(Continues)

Table A6. Replication of analysis in Table 3–5, incl. Water supply indicator baseline values

	Change in Water Consumption lpcd (1)	Change in Population With Access (2)	Change in Water Consumption lpcd (3)	Change in Population With Access (4)	Change in Water Consumption lpcd (5)	Change in Population With Access (6)
Country Characteristics						
Average GDP Per Capita[a]	0.078 (0.238)	−0.126** (0.016)				
Average Bilateral Aid Per Capita	0.006 (0.627)	−0.005 (0.610)				
Average Inflation Rate	0.000 (0.781)	0.000 (0.337)				
Conflict	0.005 (0.941)	0.026 (0.616)				
Polity Index	−0.006 (0.443)	0.008 (0.199)				
Africa Region	−0.201 (0.121)	0.056 (0.551)				
MENA Region	−0.040 (0.782)	0.176 (0.168)				
Project Management Characteristics						
Months Past Project Closure			−0.002 (0.219)	0.000 (0.820)		
Project Duration			0.002 (0.212)	0.002 (0.167)		
Second Phase[b]			0.059 (0.607)	0.025 (0.786)		
Third Phase			0.211 (0.449)	−0.250 (0.448)		
Risk Evaluation in T_0			0.034 (0.644)	−0.104 (0.107)		
Total Project Costs in USD[c]			0.010 (0.860)	0.104** (0.030)		
Partner Country Share in Total Project Costs			0.187 (0.420)	−0.170 (0.323)		
Project Design Characteristics						
Tariff Per m^3 Consumed [d]					−0.272 (0.142)	−0.320** (0.045)
Mixed System[e]					−0.141 (0.703)	0.092 (0.711)
Piped System					0.470*** (0.003)	0.000 (0.998)
Urban					0.073 (0.303)	−0.020 (0.721)
User Committee					0.348** (0.015)	0.035 (0.789)
Wastewater Component					0.116 (0.175)	0.043 (0.521)
Hygiene Promotion					0.018 (0.846)	0.158** (0.033)
Institutional Support					0.017 (0.832)	−0.077 (0.268)
Low Water Resources					0.134* (0.079)	−0.062 (0.347)

Table A6. (Continued)

	Change in Water Consumption lpcd (1)	Change in Population With Access (2)	Change in Water Consumption lpcd (3)	Change in Population With Access (4)	Change in Water Consumption lpcd (5)	Change in Population With Access (6)
Water Supply Indicator Baseline Values						
Water Consumption lpcd in T_0^a	−0.257*** (0.000)		−0.222*** (0.000)		−0.241*** (0.000)	
Population With Access in T_0^a		−0.050*** (0.000)		−0.064*** (0.000)		−0.040*** (0.000)
Constant	0.659 (0.202)	1.791*** (0.000)	0.718 (0.406)	−0.404 (0.578)	0.457** (0.039)	0.991*** (0.000)
Observations	83	87	77	80	89	89
R-squared	0.490	0.391	0.425	0.385	0.549	0.465
Adjusted R^2	0.435	0.329	0.358	0.315	0.491	0.396

Notes: standard errors are in parentheses; *** (**) (*) denotes significance at the 1 (5) (10) per cent level;

[a] logarithm;

[b] phase 1 is the left out category;

[c] logarithm;

[d] logarithm;

[e] non-piped system is the left out category.

See Table 1 for an overview of all project performance indicators and their definitions.