CENTER FOR POLICY STUDIES

Department of Public Policy

ACHIM KEMMERLING AND ALBANA REXHA

# Evaluating the Evaluators: When and How Does the World Bank Evaluate Its Projects in the Western Balkans

2015/5

WORKING PAPER SERIES

CEU Center for Policy Studies
CENTRAL EUROPEAN UNIVERSITY

CEU Department of Public Policy
CENTRAL EUROPEAN UNIVERSITY

## ABOUT THE PAPER SERIES

Working Papers reflect the on-going work of academic staff members and researchers associated with the Center for Policy Studies at Central European University and the wider CEU community. They are intended to facilitate communication between the Center and other researchers on timely issues. They are disseminated to stimulate commentary and policy discussion among an international community of scholars.

## ABOUT THE AUTHORS

Achim Kemmerling is an Associate Professor at the Department of Public Policy, and Albana Rexha was an MA student in Public Policy at Central European University in the 2012/2013 Academic Year.

## TERMS OF USE AND COPYRIGHT



CENTER FOR POLICY STUDIES
CENTRAL EUROPEAN UNIVERSIT Y
Nádor utca 9.
H–1051 Budapest, Hungary
cps@ceu.edu

# EVALUATING THE EVALUATORS: WHEN AND HOW DOES THE WORLD BANK EVALUATE ITS PROJECTS IN THE WESTERN BALKANS

Achim Kemmerling and Albana Rexha

## CONTENTS

## ABSTRACT

The World Bank is one of the leading organizations in the evaluation of development projects. It also makes results and methods more and more transparent. This paper uses the information to 'evaluate' the evaluator. We collect project evaluations of the World Bank in the Western Balkan for a period of ten years. We use this information to investigate the determinants of a) whether or not an evaluation has happened and b) what drives the results. We find that projects in poorer and less democratic countries are more likely to be evaluated. Once we control for this selection we also find that the evaluation scores for the bank and the partner institutions are correlated. We draw some normative conclusions on whether or not the World Bank should change its strategy of evaluation.

## LIST OF ACRONYMS

| | |
|---|---|
| ADB | Asian Development Bank |
| CGD | Center for Global Development |
| DAC | Development Assistance Committee |
| ECG | Evaluation Cooperation Group |
| EU | European Union |
| IMF | International Monetary Fund |
| IEG | Independent Evaluation Group |
| M&E | Monitoring and Evaluation |
| OECD | Organization for Economic Co-operation and Development |
| OED | Operations Evaluation Department |
| DFID | UK Department for International Development |
| ISR | Project Implementation Status and Results |
| ISCR | Project Implementation Status and Completion Results |

## 1. INTRODUCTION

*"In my eyes, Americans as well as other tax payers are quite ready to show more generosity. But one must convince them that their generosity will bear fruit, that there will be results."*
Paul Wolfowitz, President, World Bank

Academics and policymakers put increasing emphasis on the effectiveness of development assistance. However, the academic debate on aid effectiveness is highly controversial. On the one hand, scholars such as Easterly (2006), and Deaton (2013) argue that development assistance does have zero or even ineffective and harmful impact to the lives of poor people. On the other hand, scholars such as Sachs (2005), and Riddell (2007) argue that foreign aid has had an immense impact on reducing global poverty. Beyond the academic world, policy makers and international organizations note increasing pressure on aid budgets from tax-payers demanding accountable and effective aid (OECD 2010). Both the uncertainty surounding the effects of aid and the donors' pleads for more accountability leed to an increasing demand for evaluating development assistance.

As a leading donor agency the World Bank has championed the rise in evaluation on a massive scale and publishes more and more information on evaluation results (Cracknell 1989). The World Bank defines evaluation as "the process of determining the worth or significance of a development activity, policy or program... to determine the relevance of objectives, the efficacy of design and implementation, the efficiency or resource use, and the sustainability of results. An evaluation should (enable) the incorporation of lessons learned into the decision-making process of both partner and donor" (WB 2011).

The Bank's move towards more and better evaluation has found strong reasonance among parts of the academic community (Banerjee at al 2008, Duflo et al 2006). Yet, it also has provoked a lot of criticisms. Some economists doubt the feasibility of impact analysis for large socio-economic change (Banerjee and Duflo 2011, Easterly 2001&2006). Political economists and sociologists have criticized this impetus to quantify the unquantifiable and the technocracy of impact assessment for being both non- and highly political: non-political because it underestimates normative tradeoffs in aid; political in as much as it gives legitimacy only to certain types of projects and stakeholders but not others (e.g. Sanderson 2002).

The good news is that the available information provides ample opportunities to turn the evaluation on its 'head' and evaluate the evaluators. This can reveal important information not only about the evaluation process and its outcomes, but it can also inform a discussion whether or not to change the practice. For this purpose, we study the results of World Bank project evaluations. More specifically, we have selected over 100 World Bank projects implemented in Western Balkan[1] over the last 10 years.

We are interested in two types of questions, both of which are closely linked: a) when and why are projects evaluated (and results published); b) are there systematic trends in the results that might come from the evaluation process? Much of the process of evaluation is not accessible for third parties. Even if the World Bank tries to become more transparent and accountable, it still remains, to a certain extent, a black box. Thus, to answer these questions we employ a 'naive' reconstruction of the World

1 Albania, Bosnia and Herzegovina, Kosovo, Macedonia, Montenegro, Serbia

Bank's rationale: by looking at systematic patterns in the results we draw conclusions to the interior mechanisms of this black box.

In practical terms, we investigate the empirical determinants for the likelihood of being evaluated and the evaluation results. We find that projects are more likely evaluated if they are implemented in poorer and less democratic countries. Sound institutions, which has been one of the World Bank's prime concerns for the last decade, do seem to play less of a role. Once we control for these selection effects we find that the bank's evaluation scores and those of the receiving institutions are correlated. This implies that the evaluation process for the World Bank and for its clients is not independent from each other.

These findings, though somewhat limited in geographic scope, have interesting normative implications. The fact that poverty and democracy play more of a role than good governance in a narrow sense, lends credibility to the idea that evaluation is, first and foremost, about accountability and, ultimately, the legitimacy of the World Bank itself. Moreover, the fact that the evaluation is non-random could imply that an evaluation of the effectiveness of evaluation is limited. The fact that results of donor and recepient evaluations are correlated implies that the evaluation process may suffer from an admittedly moderate, but clearly visible bias in the evaluation technique.

To corroborate these claims we proceed in three steps. The next session reviews the debate on aid effectiveness and the rationales given for evaluation. It also discusses the challenges of evaluating development aid. The third section briefly describes the World Bank's approach to evaluation. This includes information about the Independent Expert Group (IEG), the responsible body for evaluations within the bank and outlines its procedures for evaluating projects which will also be used in data analysis. The fourth section presents and analyzes the data for the Western Balkan countries. The final section discusses the normative and positive implications of our results as well as their limits.

### 1.1. The Trend Towards Effectiveness and Accountability in Aid

The question whether aid works is as old as aid itself (Riddell 2007). Scholars hold very different views about this. Critics such as William Easterly (2006), Angus Deaton (2013) or Dambisa Moyo (2009) argue that aid has failed to show any positive impact on development in third world countries. Others scholars such as Jeffrey Sachs (2005) defend the usefullness of aid. A 'Salomonic' middle ground seems to fall in between both extremes: Scholars like Collier & Dollar (1998), Alesina & Dollar (2000) have famously argued that aid is effective in the right context, and in particular only in those countries with sound institutions and sound macroeconomic policies. From an academic point of view, this claim may rest on feet of clay as much any other finding based on macroaggregate non-experimental statistics (Easterly et al. 2004 ). And yet, this middle ground has found considerable resonance with policy makers and has given those advocating evaluation an enormous stimulous.

Two additional factors have increased the pressure of donors to evaluate their projects. First, the evaluation industry has seen its own revolution and increasing mainstreaming of using randomized experiments, due to their close collaboration between reserchers and implementers becuase they allow the estimation of certain parameters that couldn't be otherwise evaluated (Banerjee et al 2008, 2011). Thus, the move towards randomized controlled trials has given the whole evaluation industry a big push. Second, and perhaps more importantly, budget-constrained governments feel pressure from tax payers who demand higher levels accountability and transparency in the use of public money (Cracknell 2000; Hailey et al. 2005; OECD 2010).

Academics and donor agencies hold definitions of evaluation similar to the World Bank's quoted above. For example, Michael Scriven, defines evaluation as "a process of determining merit, worth, or significance" (2007). CARE defines evaluation as "the periodic assessment, analysis and use of data about a project" (Barton 1997). The Development Assistance Committee (DAC), defines evaluation as "the systematic and objective assessment of an on-going or completed development intervention, its design, implementation and results" (OECD, DAC 2010).

Despite these similarities, the purposes of evaluation are manifold and not always well aligned. A classic distinction is that evaluation has to do with proving that changes are taking place, and improving future aid interventions (Cracknell 2000). This distinction is closely related to two leitmotifs 'accountability' and 'lessons learned'. In this sense, 'accountability' means reporting your activities - the way you spend the money, to a higher authority, your principle (Crawford et al. 2003). In this respect it is interesting to note that in the accountability framework most often the ultimate principles are the tax payers in donor countries, rather than the clients of the projects in recipient countries (Easterly 2006). 'Lessons learned', on the other hand, has to do with identifying why certain activities failed and some others succeeded in reaching their objectives. In other words, it is much closer to the idea of aid effectivness.

Evaluations can also be done for formative and summative purposes. According to Scriven (1967), formative evaluations are used on the early stages of project development, while summative evaluations are conducted at the final phase of project life, in order to see the "final product". Nowadays, formative evaluation is more often referred to as monitoring of outputs, wheras summative evaluations deals with outcomes for and impacts in society (Asian Development Bank 2006). Crawford and Bryce (2003) see monitoring and evaluation (M&E) as "intimately linked", and UNDP defines M&E as two processes that "differ but are closely related" (1997).

There is considerable disagreement about the normative goals of evaluation and performance measurement. Should evaluation be focussed on the classic trinity of efficiency, efficacy, and effectiveness? Or should it take a broader stance including socio-economic impact, considerations of equity and sustainability to name but a few (Checkland 2001, Crawford et al. 2004)? Given the multitude of normative goals (and their inherent tradeoffs) it is unlikely that any evaluation will reach consensus among all stakeholders.

Closely related to the multiplicity of goals is the multiplicity of methods. For instance, impact evaluations can be of quantitative and qualitative nature (Asian Development Bank 2006, Baker 2000). However, due to the recent rise of the randomistas and the advocates of the 'management by numbers' quantitative approaches and their close kins are much more predominant than qualitative designs. The rise of the RCTs is particularly interesting also for the purposes of this paper. The experimental or randomized design works if treatment is applied randomly (Duflo et al 2003, Banerjee et al 2008, Baker 2000). Turning evaluation on its head would imply that evaluation itself should be randomized, especially in those context were it is very costly (see below).

Purposes of evaluation will, finally also differ if the evaluation is conducted by an internal or an external body. According to Weiss (1998), both types has their own advantages and disadvantages with regard to "administrative confidence", "objectivity", "understanding of the program", and "autonomy". Proximity to objectivity and autonomy make external evaluation advantageous. Where however, an intimate understanding of the program is fundamental for a successful evaluation, it might be better to recur to internal evaluation processes. In a nutshell, internal evaluators can be more practical as

opposed to external, but internal evaluators are viewed as less credential when compared to external auditors who have more reputation.

The multiplicity of purposes for evaluation is not the only problem of finding an adequate evaluation technique. There are at least three other major concerns: the involvement of politics, the high costs, and the validity of the measurement (Chelimsky 1987, Weiss 1973, Baker 2000, Cracknell 2000, Crawford et al. 2003). Starting with the first, the history of politics and evaluation is almost as old as evaluation itself (Chelimsky 1987). "A theory of evaluation must be as much a theory of political interaction as it is a theory of how to determine facts" states Lee Cronbach (1980) (quoted on Chelimsky 1987). This implies that no evaluator, and hence no evaluation can be a-political. Another way that politics gets involved in evaluations is through aid interventions providing social and economic changes which "attract" a range of stakeholders who want to see and show results to tax payers (Crawford et al. 2003). For instance, in many contexts RCTs are "politically unfeasible" (Grun 2006).

Secondly, conducting evaluation requires the use of extra financial resources (ADB 2006), thereby making it expensive for small –scale development agencies, or sometimes even for the bigger ones. This becomes a greater challenge considering due to the fact that donor agencies are mainly the ones who pay for the evaluation. The donor agencies will also influence the choice of indicators for the evaluation of a developmental activity (Bamberger 2000). A well- designed evaluation includes for professional evaluators and extensive data collection which raises the cost of evaluation (Crawford et al.2004). For these reasons, not all projects are evaluated. But if so, the decision which projects are evaluated and why becomes an important one.

The third challenge concerns a group of developmental activities that are more complicated to measure due to their less tangible processes. Projects such as capacity building tend to be described as "ill-defined" processes which present another evaluation hurdle for development community (Hailey et al. 2005; Kuehl 2009). Complexity can pose serious limits to perform an encompassing type of evaluation (Taylor 2003). It also increases the risk of meaasurement error, especially in the course of 'converting' qualitative data into quantitative scores. Among others the USAID has criticized the subjective nature of capacity building projects which can lead to "false" measurements of developmental results (USAID 2000).

To sum up, all the multiplicity of evaluation purposes as well as the practical hurldes lead to the fact that every type of evaluation will ultimately be political, partial and selective. Evaluation has become a crucial and a widely used tool in the development community but its main stimulus seems to come from the debates about aid effectiveness and aid accountability (Naonobu et al. 2009). The tax-payers want to see if their taxes are being spent efficiently, and if that is making any difference in the life of others, because that portrays the basic purpose behind all the money spend by development agencies (Riddell 2007).

### 1.2. The World Bank's Process of Evaluation

Over the last 40 years the World Bank has invested considerable time and resources on improving its evaluation techniques (Cracknell 1989; Grasso et al. 2003). The Operations Evaluation Department (OED) started in 1970 and soon gained independence by choosing its own director, and by reporting directly to the board of directors (Grasso et al. 2003). Reporting directly to the board leaves less room for influence by directors of developmental activities. In 2005 the OED became the Independent

Evaluation Group (IEG). IEG is the only independent body in WB structures in charge of evaluating developmental activities including programs, projects, and policies.

Officially, IEG's main aim is to evaluate the effectiveness of (aid) interventions (IEG 2013, Pitman et al. 2005, Tuan 2012). In particular, it assesses the "relevance", "efficiency", and "efficacy" of WB developmental activities in order to measure the contribution of WB in achieving greater developmental effectiveness. According to a recent self-evaluation, the IEG certifies its mission, purpose and independence from the World Bank group (IEG 2011, Linn 2012).

The evaluation process itself, however, is not easy to review as an outsider to IEG. The details of the evaluation process, the identities of the evaluators and the precise methodology and justification of the performance measures can only be reconstructed post hoc.

In contrast, overall transparency of evaluation results has increased considerably over time. The OECD's Development Assistance Committee (DAC) have committed themselves to strengthening the effectiveness of aid. The criteria of 1991 DAC/OECD report included effectiveness, efficiency, impact, relevance, and sustainability. Part of the report discussed the involvement of other stakeholders which represents the intended beneficiaries and also those indirectly affected by the aid intervention, in the evaluation process. In the last 20 years, the DAC criteria have spread throughout the development community (Chianca 2008).

In accordance with these donor initiatives, the World Bank publishes all its evaluations online in its Projects Database (The World Bank Group 2013). According to IEG, the WB Project Performance Rating database is the oldest database of its kind that includes about 8000 project evaluations since its establishment (WB Rating). IEG provides two types of reports for project evaluations: the Project Implementation Status and Results (ISR), and Project Implementation Status and Completion Results (ISCR).

The ISR is a standardized, four page document that includes basic information about the project such as title, country, status, sector, approval date, closing date, and commitment amount. The ISR also summarizes main objectives of the project, and offers a yes or no answer on the achievement of objectives (PISR, 2012). Another important part of the ISR are the "results indicators". Each result indicator in the ISR has the name, the "unit of measure", the "baseline", the "current", and the "end target". "Unit of measure" depends on the type of indicator chosen. For example, units could be measured in days, or it could be a simple yes or no answer. The "baseline" measures the situation before the aid intervention take places. The "current" data is measured during the project life and shows the situation by the time project is measured. The "end target" is measured after the project is finished and in a way shows if the objectives are achieved. All in all, the ISR gives a compact if somewhat stylized picture at a point in the life cycle of the project.

Compared to the ISR, the ISCR is a longer report provides more information about the project life and end results. It is not a standardized form in terms of page number or the amount of information per section, albeit it has similar sections in each evaluation report including the basic info about the project, a summary, the relevance and outcomes. Nowadays, three types performance ratings are available (WB Ratings):

- *Project outcome performance* is a rating that "captures the extent to which a project'smajor relevant objectives were achieved or are expected to be achieved, efficiently" (IEG 2011).

- *Bank performance* evaluates the quality of services the WB provides to its beneficiaries, at the beginning and throughout the project life cycle (OED 2005).
- *Borrower performance* evaluates the performance of the beneficiary. OED (2005) defines this rating as the degree to which the borrower shows willingness and ability to guarantee and comply with the criteria and requirements agreed with the bank.

The three performances are rated using the same six scales including highly satisfactory, satisfactory, moderately satisfactory, moderately unsatisfactory, unsatisfactory, and highly satisfactory, as shown below in Table 1 .

**Table 1. IEG Performance Rating Scales**

| | Scales | Explanation |
|---|---|---|
| 1 | Highly Satisfactory | *No* shortcomings in identification, preparation, or appraisal |
| 2 | Satisfactory | *Minor* shortcomings in identification, preparation, or appraisal |
| 3 | Moderately Satisfactory | *Moderate* shortcomings in identification, preparation, or appraisal |
| 4 | Moderately Unsatisfactory | *Significant* shortcomings in identification, preparation, or appraisal |
| 5 | Unsatisfactory | *Major* shortcomings in identification, preparation, or appraisal |
| 6 | Highly Unsatisfactory | *Severe* shortcomings in identification, preparation, or appraisal |

Source: Author's compilation based on 2005 OED report

## 2. DATA ANALYSIS AND RESULTS

We collected data from the World Bank dataset on "Projects and Operations" on a total of 109 implemented projects in the Western Balkan for a period of ten years. The information includes the project name, country, approval date, closing date, commitment amount, % spend in each sector for each project (usually a project is spread into several sectors), intermediate results indicator which are conducted while the project is running, implementation completion results report which is conducted at the end phase of the project cycle, project outcome performance, bank and borrower performance.

The projects implemented show a considerable variety over sectors and the size ranges from 0.1 to 111 million $. Table 2 shows the descriptive statistics of our data. We see for instance, that the project amounts of evaluated an non-evaluated differ from country to country. For our purposes it is most interesting to note that not all projects were evaluated (or the evaluation published): only 68 out of 109 projects have an evaluation report. The shares range from only every second project in Serbia to more than 80 percent in Macedonia.

**Table 2. Descriptive Statistics**

| Country | Albania | B & H | Kosovo | Macedonia | Montenegro | Serbia | Total |
|---|---|---|---|---|---|---|---|
| Evaluated? | 0.67 | 0.45 | 0.67 | 0.81 | 0.7 | 0.54 | 0.62 |
| Outcome Performance | 3.31 | 3.33 | 3.6 | 3.31 | 3.14 | 3.85 | 3.44 |
| Borrower Performance | 3.06 | 3.63 | 3.44 | 3.18 | 3.8 | 3.75 | 3.41 |
| Bank Performance | 3.31 | 3.5 | 3.56 | 3.42 | 3.5 | 3.67 | 3.48 |
| Amount | 10.86 | 18.39 | 4.88 | 19.63 | 14.9 | 37.3 | 18.9 |
| GDP p.c. | 2.76 | 3.18 | 2.49 | 3.34 | 4.88 | 4.1 | 3.37 |
| Poverty Headcount | 17.83 | 15.33 | 36.78 | 19.3 | 8.33 | 9.62 | 17.81 |
| V & A | 0.06 | 0.12 | -0.39 | 0.05 | 0.18 | 0.04 | 0.02 |
| Government Effectiveness | -0.48 | -0.71 | -0.34 | -0.17 | 0.07 | -0.27 | -0.38 |
| RoL | -0.74 | -0.52 | -0.87 | -0.4 | -0.15 | -0.68 | -0.6 |

Source: Authors compilation based on the dataset created based on World Bank data-base

This prompts the question what makes a project being evaluated or not. We should expect that several factors make projectes more likely to be (publically) evaluated: First, higher salience, measured by higher project sums, should increase the likelihood of evaluation. Second, since it is the World Bank's mission to eradicate poverty, it might be especially sensitive to context of high poverty. Hence we should think that projects in countries with higher poverty rates or lower GDP per capita are more likely to be evaluated. Third, since the World Bank seems to be believe that effectiveness depends on the quality of domestic institutions and policies (see above), we would also expect that projects in countries with lower levels of good governance should be more likely scrutinized. Fourth, if, however, the World Bank is more concerned about accountability and ultimately its own legitimacy towards its principles we should expect that political indicators such as democracy should be more relevant.

To reconstruct the World Bank's rationale we performed some simple comparisons of means tests between the two subgroups: projects evaluated, projects not evaluated. Table 3 shows the results. Going through the list of variables, we find that project sum (amount committed) does not seem to be markedly different for the two subgroups. For both groups the average is around 19$ million. The mean for those evaluated is some two millions higher, but the difference is not large enough to justify systematic tendencies in the data. Indicators of wealth and poverty, however, do suggest some effect. GDP per capita of the country is some 850$ lower in those projects that were evaluated. The difference is one fourth of the average GDP p.c. in the region and large enough to suggest a systematic pattern. Similar things apply to poverty measures, in this case the poverty headcount (World Bank Data). Evaluated projects are in countries with a poverty headcount of 1.5 times of the poverty headcount of countries whose projects were not evaluated.

As for the indicators of good governance it is interesting to note that the strongest results seem to come from Voice & Accountability (V&A). The World Bank defines V&A as "measuring perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association and a free media" (Kaufmann et al, 2008; Thomas 2010 ). Table 3 suggests that projects evaluated tend to be more often in countries with weaker indices of V&A then projects not evaluated. In turn, Government Effectiveness (GE), shows no systematic effect. GE is defined as "measuring the quality of public services, the quality of the civil service and the degree

of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies." This is remarkable as one would expect that if the World Bank prioritizes effectiveness, government effectiveness should be an important context factor. The results, however, seems to suggest that the purpose of evaluation is more related to accountability of the World Bank in political terms, than to aid effectiveness in a narrow sense.

As for the third indicator, Rule of Law (RoL) we do find some systematic differences. As expected, projects are more likely to be evaluated in context where RoL is weak. RoL is defined as "measuring perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, the police and the courts, as well as the likelihood of crime and violence." This finding seems to imply that the World Bank is also concerned about aid effectivess in highly unstable contexts, but the effect is much less visible than the one for V&A.

**Table 3. T-Tests Comparing Evaluated and Non-Evaluated Projects**

| | mean | difference | sign. | std. err. | nobs. |
|---|---|---|---|---|---|
| **Amount** | 18.9 | 2.12 | | 5.36 | 109 |
| **Gdp p.c.** | 3372.55 | -849.54 | *** | 246.32 | 109 |
| **Poverty headcount** | 17.81 | 8.05 | *** | 2.51 | 53 |
| **V&A** | 0.015 | -0.11 | *** | 0.04 | 109 |
| **Gov. Effectiveness** | -0.38 | 0.01 | | 0.05 | 98 |
| **RoL** | -0.60 | -0.09 | * | 0.05 | 109 |

Levels of Significance: * < .1; ** <.05; *** <.01

To conclude, our somewhat naive reconstruction of the World Bank's motives implies that project evaluation is not completely random, but focusses on projects in poorer countries. It also seems to be strongest related to issues of V&A and hence more consistent with the idea that the World Bank does it for purposes of self-legitimization than for matters of aid effectiveness. This is important information in itself (see below), but the reasons for selecting projects for evaluation might also loom large into the actual performance measures of the evaluated projects. This is what we turn to now.

This section assess the 68 projects that have evaluation reports with detailed information as opposed to non-evaluated projects which have only specific information available such as commitment amount. Apart from general data including country, commited amount, the information found in the evaluation reports consists also of data such as outcome performance, bank performance, borrower performance. Table 2 showed the country averages range between satisfactory and highly satisfactory. Some countries such as Serbia have a higher average than others like Montenegro. However, as the previous section has made clear, this may be due to selection effects. What we don't see is how non-evaluated projects would have fared. Are Serbian projects really better, or do we just not see the other half of the Serbian projects?

A good start into these questions is by merely looking at simple pairwise correlations between these three performance indices. Table 4 shows the results.

**Table 4. Correlations between Performance Indices**

| | Outcome Performance | Bank Performance | Borrower Performance |
|---|---|---|---|
| **Outcome Performance** | 1 | | |
| **Bank Performance** | 0.29**<br>(0.04)<br>49 | 1 | |
| **Borrower Performance** | 0.64***<br>(0.00)<br>61 | 0.10<br>(0.51)<br>46 | 1 |

Levels of Significance: * < .1; ** <.05; *** <.01

The results show that the overall performance primarily depends on the borrower's performance, but also on the bank's performance. This is logical and to be expected, since the overall performance should be closely related to the evaluation of the two partners involved. We also see that the borrower's performance and the bank's performance are not correlated. This is a good sign of an evaluation method. Correlation between bank's and borrower's performance would imply that both parts might bias each other and that the evaluation outcome is hard to interpret. The fact that we do not find a correlation lends credibility to the claim that the evaluation is independent and neutral.

However, the previous section has shown that the selection of who is evaluated or not is not random. It is therefore crucial to see whether the selection has larger implications also for the results of the evaluation process. For this purpose we run some statistical analyses that predict the performance indicator with the help of other variables. The first step lies in converting qualitative assessments into numeric scores, ranging from the lowest (1) 'unsatisfactory' to the highest (5) 'highly satisfactory'. A look at the descriptives in table 2 shows that the regional averages performance values are very similar to each other (ranging between 3.41 and 3.48). There does not seem to be any systematic tendency against either borrower or the bank.

The next step is to run some simple regressions with the performance measures as dependent variables.[2]

2 The data is essentially ordinal in nature so we should run ordinal regressions. However, the number of observations is not very large which poses a problem for methods based on maximum likelihood. Moreover the results essentially confirm the simpler analysis so we decided not to report them. They are available on request.

**Table 5. Bank and Borrowers' performance**

| | (1) Bank Value | (2) Bank Value | (3) Bank Value | (4) Borrower Value | (5) Borrower Value | (6) Borrower Value |
|---|---|---|---|---|---|---|
| **Poverty** | -0.000438 | -0.0175 | -0.0169 | -0.0167 | 0.00308 | 0.00139 |
| | [0.023] | [0.026] | [0.020] | [0.035] | [0.042] | [0.038] |
| **GDP p. c.** | 0.424*** | 0.0235 | | 0.136 | -0.0756 | |
| | [0.172] | [0.305] | | [0.143] | [0.492] | |
| **V& A** | -1.272 | -1.626* | -1.517 | -0.972 | 0.890 | 0.633 |
| | [0.875] | [0.885] | [1.017] | [1.347] | [1.542] | [1.852] |
| **Borrower Value** | | 0.269* | 0.269*** | | | |
| | | [0.132] | [0.116] | | | |
| **Bank Value** | | | | | 0.701* | 0.705*** |
| | | | | | [0.343] | [0.304] |
| **Constant** | 2.015*** | 2.433* | 2.527*** | 3.269*** | 1.445 | 1.190 |
| | [0.782] | [1.304] | [0.531] | [0.925] | [2.274] | [1.185] |
| **Selection Stage** | | | | | | |
| **GDP p. c.** | | | -1.434*** | | | -1.697*** |
| | | | [0.601] | | | [0.650] |
| **V& A** | | | -4.817*** | | | -9.559*** |
| | | | [2.443] | | | [4.305] |
| **Poverty** | | | -0.136*** | | | -0.260*** |
| | | | [0.069] | | | [0.120] |
| **Constant** | | | 6.982*** | | | 10.16*** |
| | | | [2.644] | | | [3.726] |
| **athrho** | | | -0.192 | | | 0.192 |
| **_cons** | | | [1.064] | | | [0.829] |
| **lnsigma** | | | -1.008*** | | | -0.531*** |
| **_cons** | | | [0.165] | | | [0.153] |
| **N** | 29 | 23 | 47 | 34 | 23 | 42 |
| **R2** | 0.225 | 0.379 | | 0.053 | 0.191 | |

Standard errors in brackets

* $p < 0.10$, *** $p < 0.05$, *** $p < 0.001$

The table 5 shows six different regressions. The first three are for the bank's performance and the other three are for the borrower's performance as the dependent variables. In each case, the first model is the baseline and includes the (statistically significant) variables of the selection stage: GDP per capita, poverty and V&A. These variables are meant to capture the non-randomness of the evaluation. The second (and fifth) model include the other performance measure to check for the covariation between the bank's and the borrower's evaluation. Models (3) and (6) perform the same analysis, but in a Heckman selection framework. The Heckman model explicitly captures the selection bias which might arise due to the non-randomness of the data (Heckman 1979).

Let us start with models (1) and (4). We see that the bank's performance seems to be better in countries with higher GDP per capita. The borrower's performance shows no such results. Poverty and V&A don't seem to affect the results very much. However, if we plug in the value's for the borrower (model (2)) and the bank (model (5)), we do find that they are positively correlated with each other. This is an interesting contrast to the previous simple correlations where we did not find any effect. It seems that, by controlling for other determinants such as GDP p.c. or V&A, we do find a systematic relationship between both performance measures. The last models (3) and (6) do repeat this finding: the bank's and the borrower's performance co-vary. However, we don't find direct evidence for selection bias, i.e. the non-randomness of the evaluation does not affect the performance data directly. This essentially means that both types of regression in the Heckman model, those for the selection stage, and those for the performance stage can be run independently from each other. The lower half of the Heckman models shows this 'selection stage', i.e. whether or not a project has been evaluated or not. Essentially it repeats the findings of the simpler t-tests of Table 3.

All in all, the findings raise some concerns about the World Bank's approach to evaluation. Apparently, after controlling for mainly country-specific factors we do find a co-variation between the bank's and the borrower's evaluation. But if so, how could we tell who is responsible for the final outcome, the borrower, the Bank, or even the biased evaluator? This is an interesting point for the discussion. Moreover, if the selection of projects evaluated or not is non-random, how sure can we be about claims like the one between good governance and aid effectiveness? Among others, Isham et al. (1997) claim that development projects have better return on investment in countries with better governance and accountability. How sure are we that this is true, if evaluation is not present everywhere?

As further checks we also disaggregated the projects into sectors such as economic, law, trade, industry, telecommunication, mining, agriculture, energy, security, public administration, health, central administration, labor, social welfare, and the like. Table 6 shows the number of projects and the total project sums for each of the main sectors.[3]

**Table 6. World Bank Projects Differentiated by Sector**

| | Mining & Energy | Health/ Labor/ Social Welfare | Edu-cation | Economy/ Finance/ Banking | Public Admin-istration | Justice/ Law/ Security | Agri-culture/ Forestry/ Rural | Infra-structure/ Telecom-munication | Trade & Industry |
|---|---|---|---|---|---|---|---|---|---|
| Total Number of Projects | 14 | 31 | 7 | 9 | 51 | 8 | 12 | 10 | 14 |
| Share evaluated | 14 | 31 | 7 | 9 | 51 | 8 | 12 | 10 | 14 |
| Total Amount | 293.75 | 638.7 | 69.5 | 354.21 | 886.39 | 183.41 | 139.02 | 239.75 | 389.3 |
| Average of Amount | 20.98 | 20.60 | 9.93 | 39.36 | 17.38 | 22.93 | 11.59 | 23.98 | 27.81 |

3 Some projects span several sectors so we weighted them according to the relative share of project money that goes into each sector

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Average Evaluation Bank** | 2.79 | 3.45 | 3.14 | 3.11 | 3.12 | 3.63 | 3.25 | 3.40 | 3.57 |
| **Average Evaluation Borrower** | 2.64 | 3.39 | 3.57 | 3.00 | 3.02 | 3.63 | 3.08 | 3.10 | 3.57 |
| **Average Evaluation Outcome** | 3.57 | 3.55 | 3.57 | 3.78 | 3.02 | 3.63 | 3.58 | 3.30 | 3.79 |

We see that two sectors stand out and combine almost 50% of all projects: 'public administration' and 'health, social and labour issues'. When it comes to amounts spent the differences to other sectors is less obvious, as 'economy, finance and banking' or 'minining and energy' also received considerable amounts. The best ratings, however, are found in economy & banking and trade & industry. The worst in mining & energy. Hence, the sectoral differences, are not huge, but clearly visible. While the numbers are too small to draw any statistical inferences, the sectoral differences may also constitute a source of bias in the evaluation technique.

## 3. POSITIVE AND NORMATIVE DISCUSSION

Our paper asked two types of questions: When and why are evaluations done, and what are the results? Hence we have found two types of results as answers to these questions: those about the determinants of what makes a project being evaluated, and what makes a project being highly or lowly rated.

As for the first type of results we found that evaluation seems to be more about accountability than effectiveness. To be clear, the results are not beyond any doubt. For instance, future research should check whether similar things apply to projects in other regions. Yet the finding is interesting and worth further inspection. Is accountability the main reason and what role does effectiveness really play? Many critical scholars like Moran, Rein and Goodin (2006) would argue it is more about the former, and more about the World Bank's legitimacy than actual learning exercises. Now, accountability is no bad thing at all, but if it does not lead to other positive results it is somewhat 'toothless'.

The non-randomness also precludes, to some extent, the evaluation of evaluations. If the evaluation is not random, it is much harder to identify the effects of evaluations. For instance, it is much harder to find out whether countries/ clients improve their performance because they fear the negative impact of evaluation or not. The non-randomness makes it also much more difficult to control the quality of the evaluations (Duflo et al 2003, Banerjee et al 2008). Eventually, these problems carry over to the actual results of evaluations. When it comes to performance measurement, there are numerous and well-known problems of quantifying the 'unquantifiable'. However, our results show that even within the technocratic vision of the World Bank it might be difficult to avoid bias between the assessment of the bank and the borrower.

Again our results are somewhat preliminary and should be substantiated with more sophisticated analysis, better and more data. Yet they show how difficult a 'neutral' evaluation process is in practice.

This is all the more a problem, if the organization implementing the evaluation is not completely independent from the World Bank (Weiss 1998). Of course, it is difficult to perform an anonymous peer-review procedure with something as prominent as a World Bank project. However, there are solutions: One is to divulge even more information than as of now about the evaluation process. To make a concrete example: it would be very helpful if the names and biographies of the evaluators were known in each case. This is very helpful information, because one would be able to reconstruct 'the priors' of individuals as well as their relationship to the Bank or the borrower. In general, making the methodology and the process of the evaluations more transparent would not only help to increase accountability but also effectiveness of the evaluation process.

All in all, the World Bank has come a long way in the way how it deals with performance, evaluation and transparency. Given the openness one can only hope that this will stimulate further steps towards an even more open and transparent evaluation procedure in the future.

## BIBLIOGRAPHY

Asian Development Bank (2006). Impact Evaluation: Methodological and Operational Issues. *Published in Philiphines*.

Banerjee, A. And Duflo,E. (2008). The Experimental Approach to Development Economics. Working Paper. The National Bureau of Economic Research Publishing.

Banerjee, A. And Duflo, E. (2011). Poor Economics: A Radical Rethinking of the Way We Fight Global Poverty. Public Affairs.

Barton, T. (1997). Guidelines to Monitoring and Evaluation. How are we doing? *CARE International, Uganda*.

Burnside, A.C. and D. Dollar (2000). Aid, policies, and growth. *American Economic Review.*

Baker, J. (2000). Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners. *Washington: World Bank.*

Bamberger, M. (2000). The Evaluation of International Development Programs: A View from the Front. *American Journal of Evaluation*.

Bamberger, M. (2009). Institutionalizing Impact Evaluation Within the Framework of a Monitoring and Evaluation System. Poverty Analysis, Monitoring, and Impact Evaluation Thematic Group. *The World Bank.*

Collier, P. and D. Dollar (1998). Aid Allocation and Poverty Reduction. *Washington DC: World Bank*

Cracknell, B. (2000). Evaluating Development Aid: Issues, Problems, and Solutions. *Sage Publications: New Delhi.*

Crawford, P., Bryce,P. (2003). Project monitoring and evaluation: a method for enhancing the efficiency and effectiveness of aid project implementation. *International Journal of Project Management.*

Checkland, P. B. (2001). Soft Systems Methodology, in J. Rosenhead and J. Mingers (eds) *Rational Analysis for a Problematic World Revisited. Brisbane: Wiley*

Crawford, P., Perryman, J., Petocz, P.(2004). Synthetic Indices: A Methods for Evaluating Aid Project Effectiveness. *Sage Publication.*

Chelimsky, E. (1987). The Politics of Doing Program Evaluation. *Society, Springer, New York*.

Chianca, T. (2008). The OECD/DAC Criteria for International Development Evaluations: An Assessment and Ideas for Improvement. *Journal of Multidisciplinary Evaluation*.

Deaton, A. (2013). The Great Escape: Health, Wealth, and the Origins of Inequality. *Princeton University Publishing.*

Duflo, E. and Glennerster, R. and Kremer, M. (2006). Using Randomization in Development Economics Research: A Tooklit. The National Bureau of Economic Research Publishing.

Duflo, E and Kremer, M. (2003). Use of Randomization in the Evaluation of Development Effectiveness. World Bank Operations Evaluation Department. Conference on Evaluaton and Development Effectiveness in Washington, D.C.

Easterly, W. And Ross, L. and David, R. (2004). Aid, Policies, and Growth: Comment. American Economic Review

Easterly, W. (2006). The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good. *New York: Penguin Press*.

Easterly, W. (2007). Are Aid Agencies Improving? *Economic Policy*.

Easterly, W. (2001) The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics. *Cambridge, MIT*.

Freedom of House (2012).Rating and Democracy Score, Nations in Transit. http://www.freedomhouse.org/sites/default/files/2012%20%20NIT%20Tables.pdf

Grun, R.(2006). Monitoring and Evaluating Projects: A step by step Primer on Monitoring, Benchmarking, and Impact Evaluation. *Health, Nutrition, and Population (HNP), World Bank publishing.*

Grasso, P., Wasty, S., Weaving, R. (2003). World Bank Operations Evaluation Department. The First 30 Years. *WB publishing, Washington DC*.

Hailey, J., James, R. Wrigley, R. (2005). Rising to the Challenges: Assessing the Impacts of Organizational Capacity Building. *Praxis Paper No.2 : The International NGO Training and Research Center*.

Heckman, J. (1979). "Sample Selection Bias as a Specification Error." *Econometrica, 47:1*, pp. 153-62.

Independent Evaluation Group (2011). Results and Performance of the World Bank Group. http://ieg.worldbankgroup.org/content/ieg/en/home/reports/rap2011.html

Independent Evaluation Group (2013). Improving World Bank Group Development Results through Excellence in Evaluation. About IEG. http://ieg.worldbankgroup.org/content/ieg/en/home/about.html

Independent Evaluation Group (2011). Access to Information Policy. IEG: WB,IFC, MIGA publishing.

Independent Evaluation Group. Access to Information. http://ieg.worldbankgroup.org/content/ieg/en/home/access_to_information.html

Independent Evaluation Group (2013). Evaluation Week. http://ieg.worldbankgroup.org/content/ieg/en/home/Events/evaluationweek.html

Independent Evaluation Group (2011). Self-Evaluation of the Independent Evaluation Group. World Bank Publishing, Washington, D.C.

Independent Evaluation Group (2013). World Bank Project Performance Ratings. http://ieg.worldbankgroup.org/content/ieg/en/home/ratings.html

Isham, J. Kufmann, D. Pritchett, L. (1997). Civil Liberties, Democracy, and the Performance of Government Projects. *Oxford Journal, The World Bank Economic Recovery*.

Jeffrey, S. (2005). The End of Poverty: Economic Possibilities for Our Time. *New York: Penguin Press*.

Kaufmann, D. Kraay, A. Mastruzzi, M. (2008) Governance matters VII: Aggregate and individual governance indicators 1996–2007. *Washington DC: World Bank*.

Kuhle, S. (2009). Capacity Development as the Model for Development Aid Organizations. *International Institute of Social Studies, The Hague.*

Labs, E. (1997). The Role of Foreign Aid in Development. T*he University of Michigan.*

Linn, Jonathan (20120). Evaluating the Evaluators: Some Lessons from a Recent World Bank Self-Evaluation, Brookings Opinions, http://www.brookings.edu/research/opinions/2012/02/21-world-bank-evaluation-linn.

M. Moran, M. Rein & R. E. Goodin (2006). The Politics of Policy Evaluation. In Oxford Handbook of Public Policy (pp. 317-335). *Oxford: Oxford University Press*.

Moyo, D. (2009). Dead Aid. Why Aid is Not Working and How There is Another Way for Africa. *Penguin Books.*

Mohr, L.B. (1995). Impact Analysis for Program Evaluation. 2nd ed. Thousand Oaks, Calif.: *Sage Publications*.

Marinaccio, M. and Trojanowski, M. (2012). Projects, Programs Defined. *Internal Auditor. Illustration by Sean Yates/ Yacinsky Design LLC*.

Naonobu, M. and Nobuko, F. (2009). Evaluating Development Assistance: A Japanese Perspective. *Foundation for Advances Studies on International Development, International Development Research Institute, Tokyo, Japan*.

OECD (2010). Evaluation in Development Agencies, Better Aid. *OECD Publishing.*

OECD, DAC Network on Development Evaluation (2010). Quality Standards for Development Evaluation. *OECD Publishing*.

OECD (1991). Principles for Evaluation of Development Assistance. *Development Assistance Committee, Paris*.

OECD (2010). Quality Standards for Development Evaluation. DAC Guidelines and References Series. *OECD Publishing*.

Operations Evaluation Department (2005). Harmonized Evaluation Criteria for ICR and OED Evaluations. *WB Publishing, Washington, D.C.*

Pacquement, F. (2010). How Development Assistance from France and the United Kingdom Has Evovled: Fifty Years from Decolonisation. *Geneva-Graduate Institute of International and Development Studies*.

Palumbo, D., Nachmias, D. (1983). The Preconditions for Successful Evaluation: Is There an Ideal Paradigm? *Policy Science: Elsevier Science Publisher B.V., Amesterdam.*

Project Implementation and Status Results (PISR) (2012). http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/ECA/2012/06/27/C1413BC47957F75985257A2A006FBB10/1_0/Rendered/PDF/ISR0Disclosabl027201201340828423956.pdf

Pitman, G., Feinstein, O., Ingram, G. (2005). Evaluating Development Effectiveness. World Bank Series on Evaluation and Development. *Volume 7, Transaction Publishers. New Brunswick and London.*

Riddell, R. (2007). Does Foreign Aid Really Work? *Oxford University Press*.

Sanderson, I. (2002). Evaluation, Policy Learning, and Evidence-Based Policy Making, Public Administration, 80/1, pp. 1-22.

Scriven, M. (2007). The Logic of Evaluations. *Department of Psychology, Claremont Graduate University. USA*.

Scriven, M. (1967). The Methodology of Evaluation. In Perspectives of curriculum evaluation, eds. Ralph W. Tyler, Robert M. George. *AERA Monograph Series on Curriculum Evaluation, vol. 1. Chicago: Rand McNally*.

Skoufias, E., Parker, S. (2001). Conditional Cash Transfers and Their Impact on Child. *Discussion Paper, International Food Policy Research Institute, Washington, D.C.*

Taylor, J. (2003) Using Measurement Developmentally. *Cape Town: CDRA*.

Thomas,M.A (2010).What Do the Worldwide Governance Indicators Measure? E*uropean Journal of Development Research*.

Tuan, M. (2012). External Evaluation Advisory Committee Scoping Project: Findings and Recommendations. *The David and Lucile Packard Foundation.*

Transparency International (2012). Corruption Perception Index http://cpi.transparency.org/cpi2012/results/

The Center for Global Development (2006). When Will We Ever Learn? Improving Lives through Impact Evaluation. *Evaluation Gap Working Group. Washington, D.C.*

UNDP (1997). Results-oriented Monitoring and Evaluation. *UNDP publishing*.

USAID (2000). Measuring Institutional Capacity. Recent Practices in Monitoring and Evaluation Tips, Number 15. *Washington, DC. Center for Development Information and Evaluation.*

Weiss, C. (1998). Evaluation. 2nd ed. Prentice Hall, Upper Sadler River, New Yersey. Weiss, C. (1973). Where Politics and Research Evaluation Meet. *American Journal of Evaluation.*

World Bank, IPDET, Presentation. http://www.worldbank.org/oed/ipdet/presentation/M_01- Pr.pdf

World Bank. Archives. http://web.worldbank.org/WBSITE/EXTERNAL/EXTABOUTUS/EXTARCHIVES/0,,contentMDK:23164816~pagePK:36726~piPK:36092~theSitePK:29506,00.html

Work and Schooling: Evidence from the PROGRESA Program in Mexico. *Project Muse. Published by Brookings Institute Press*.

World Bank (2013). Home, Who we are, Mission. http://web.worldbank.org/WBSITE/EXTERNAL/EXTABOUTUS/0,,contentMDK:20046292~menuPK:1696892~pagePK:51123644~piPK:329829~theSitePK:29708,00.html

World Bank (2012). Doing Business Report. http://www.doingbusiness.org/rankings

World Bank (2004). Monitoring & Evaluation: Some Tools, Methods & Approaches. *World Bank Operations Evaluation Department, Evaluation Capacity Development. Washington DC.*

World Bank (2006). Doing Impact Evaluation: Impact Evaluation and the Project Cycle. *World Bank Publishing*. World Bank (2013). Projects and Operations Dataset. http://www.worldbank.org/projects