WPS7181

Policy Research Working Paper 7181

# Correlates of Success in World Bank Development Policy Lending

*Peter Moll*
*Patricia Geli*
*Pablo Saavedra*

WORLD BANK GROUP
Operations Policy and Quality Department (OPCS)
Country Economics Unit
January 2015

Policy Research Working Paper 7181

## Abstract

This paper examines the correlates of success of development policy lending operations of the World Bank between 2004 and 2012. The paper uses a data set constructed of individual loan characteristics and ex-post loan ratings produced by the World Bank's Independent Evaluation Group. Departing from the related literature, the paper focuses mostly on examining the impact of loan characteristics, reform program design features, and task team leader skills, among other variables, on intended development results, while still controlling for country characteristics. It finds that a variable used to reflect congruence or "line of sight" between the policy reforms supported and the results framework is a critical ingredient for success. Task team leader skills in general, and task team leadership by staff affiliated with the former "Economic Policy" department of the World Bank, also increase the chance of success. Conversely, a weaker set of supported reforms in these operations tends to reduce the chance of success. Reforms supported in the energy sector seem to reduce the likelihood of success, perhaps because of the inherent political difficulties of implementing reforms in this sector. The paper also draws important policy and institutional implications from these and other findings.

This paper is a product of the Country Economics Unit, Operations Policy and Quality Department (OPCS). It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at pmoll@worldbank.org.

*The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.*

Produced by the Research Support Team

# Correlates of Success in World Bank Development Policy Lending

Peter Moll, Patricia Geli, and Pablo Saavedra

Keywords: World Bank, development policy lending, aid effectiveness

JEL classification codes: O10, O19, O20, O57

## 1. Introduction

Taxpayers across countries – whether net providers or net recipients of official development financing – have been demanding that development financing should deliver tangible results. This concern has generated an increasing body of literature on the effects of aid and development financing on economic growth, poverty, inequality, and a myriad of sector specific loan outputs and outcomes. The bulk of the lending of multilateral development banks has been through investment projects (e.g., in the case of the World Bank, this has historically accounted for roughly 70 percent of total commitments). The balance of lending is in the form of policy-based, fast-disbursing loans – often termed budget support in the literature, hereafter called policy-based lending. In the case of the World Bank, these were called "structural adjustment loans" up to August 2004 and "development policy lending" subsequently.[1] The nature of this financing is different to that of investment project financing and thus it requires a separate examination.

The findings in the literature of the impact of policy-based lending are diverse, varying from those that are highly positive about the impact of this type of lending on economic and development outcomes to those that conclude that impact is limited or subject to various caveats. Much of the literature is limited to exploring the effect of country aggregate indicators (e.g., level of income, growth, macroeconomic conditions), either as core explanatory variables or as dependent variables. Moreover, most of the existing research either examines the old structural adjustment loans alone or combines the old structural adjustment loans (up to 2004) with the new development policy lending (post-2004) in a single basket, despite the significant reforms of the latter instrument in late 2004 at the World Bank (see below).

[1] Development policy lending is a fast-disbursing form of budget support, which historically accounts for around 30 percent of new commitments of the World Bank. This financing instrument is meant to support policy and institutional reforms, which in turn would render development results over the short to medium terms. All the "prior actions" or conditions for the loan are met before the approval of the World Bank Board of Executive Directors. There is no earmarking; thus the funds become part of the government's budget. From 2014 the instrument was renamed "development policy financing".

This paper examines the correlates of success of the World Bank's development policy lending in achieving its intended development results, focusing mostly on loan design characteristics while controlling for country factors. The paper examines these correlates only for the post-2004 development policy lending (hereafter policy-based loans).

To do so, the paper applies straightforward regression models on a data set that compiles performance of World Bank policy-based loans, loan design characteristics as well as country factors, between September 2004 and June 2012. The set of loan design variables constructed is based on a thorough desk review of each loan[2] Program Document and on the basis of task team leader skills and past performance.

To place this study in context, it is important to highlight the changes that occurred in 2004 to the policy-based lending instrument of the World Bank, which included the following: (a) the conditions for loans were generally to be drawn from the governments' own development strategies; (b) the conditions were to be reduced in number, and were to become significantly more selective; (c) no loans were to be granted to countries that did not have an adequate macroeconomic policy framework; and (d) there had to be more focus on development results as a consequence of the policies supported. A further important change in 2004 was the introduction of the *programmatic series of loans*.

There is a marked difference in performance between the old loans and the new, as is shown by the bottom line ratings of Implementation Completion Reviews (a review that all World Bank loans undergo after they close) and the validations of the latter by the Independent Evaluation Group (IEG) of the institution. As Dollar and Svensson (2000) report, in the structural adjustment era 36 percent of fast-disbursing loans failed; the failure rate of development policy loans in our data set, which extends from September 2004 to June 2012, was 19.6 percent using the same measure[3]. Hence the success rate of the "new" policy-based loans is high, and is higher than that of its predecessor.

[2] This paper uses "loans" as a generic term for loans, credits (from the heavily subsidized IDA window), and grants.

[3] By "failure" is meant a rating by the Independent Evaluation Group (IEG) of "moderately unsatisfactory" or poorer; "success" is a rating of "moderately satisfactory " or better.

Nonetheless the findings of this paper emphasize opportunities to keep up the performance, particularly as reform complexity may increase under the second, third and fourth waves of reforms being undertaking in emerging and developing economies.

## 2. Related literature

Part of the broader body of literature on the impact of aid and development financing on economic variables makes use of cross-country data and aid/development financing at a very aggregate level. The findings vary significantly. Burnside and Dollar's (2000, 2004) influential work finds that aid has a positive impact on growth in developing countries with good fiscal, monetary and trade policies, but has little effect when economic policies are inadequate. Clemens *et al.* (2004) find that increases in aid have been followed, with a lag, by modest increases in investment and growth. Yet, Boone (1995) finds that aid has no impact on investment, growth, or human development indicators; but it does increase the size of government. Factors such as the complexity of each country situation, the non-differentiation of the type of financing (e.g., infrastructure investments vis a vis policy-based lending), problems of endogeneity, sample size, missing data, among others, may have contributed to the variation in findings. Smets and Knack (2014) examine the impact of World Bank fast-disbursing lending on the quality of economic policy. They use elements of the Bank's Country Policy and Institutional Assessment ratings as the dependent variable. Noting that past studies have found only weak effects of fast-disbursing lending on macro stability, they restrict their sample to consider the impact of loans that were focused on macro and fiscal issues. Their key result is that fast-disbursing lending does succeed at improving governments' economic policy-making.

A second strand of inquiry is the case study approach, commonly used by multilateral banks and bilateral aid organizations for self-evaluation. These studies have the advantage of being able to focus sharply on the quality of the aid delivered within a specific country and sector setting. However, they also bear the disadvantage of a lack of representativeness, so that it is difficult to infer broader lessons for other countries – owing to the necessarily small sample sizes. The "Joint Evaluation of Budget Support" (1994-2004), for example, studied the delivery of joint policy-based lending to seven countries in great detail. It found that this modality of lending (viz. that supported by more than one national or multilateral institution) reinforced pre-existing macroeconomic stability and it supported fiscal discipline, without hampering tax effort.

A third line of inquiry focuses on project/program level outcomes and seeks to uncover factors that increase the likelihood of success in achieving the intended objectives and outcomes. Independent variables are typically country characteristics, task team leader skills, and project design. An early example is the work of Kilby (1995, 2000) which examines the effects of supervision on the ratings of World Bank-funded investment projects from 1981 to 1991, finding a positive impact of early supervision on performance. Chauvet *et al.* (2010) use data from World Bank-funded loans (mixing together investment loans and old structural adjustment and new policy-based loans) from 1977 to 2002. They similarly find that supervision is one of the factors that raise the chance of success, and that projects in private sector development are more likely to fail than those in urban development or transport. They also find that the probability of success of projects in post-conflict situations increases the longer the peace holds. Using data between 1990 and 2002 on World Bank investment projects and structural adjustment loans together, Wane (2004) finds that capable and accountable governments take on projects with better design, but low capacity and less democratic governments tend to accept lower-quality projects because they lack the ability to screen them well and because they are driven by the urgency to secure the financing regardless of the result.

Dollar and Svensson (1998, 2000) focus on 179 structural adjustment loans between 1980 and 1995. They posit a binary probit as their main model, with country-level political economy variables, variables relating to the Bank's project preparation and supervision, and some variables about the operation. They hypothesize that the preparation and supervision effort would influence the probability of success, but note that these are endogenous: if the task team leader suspects impending failure, s/he may increase the effort level. One of their key findings is that when the Bank's preparation and supervision are treated as endogenous, neither has any statistically significant effect on the probability of success. Their explanation is that *country* variables such as democratic elections are a powerful determinant of outcomes, and so if the Bank makes a bad choice of country, its adjustment programs are fated to fail, irrespective of its efforts in preparation and supervision. They imply that the best way to improve programs would be to select more promising countries in the first place. One drawback to their approach, however, is that there were no variables about the innards of the operations: the strength or relevance of the conditions, the kinds of results expected, and other aspects of the project or program design.

Denizer *et al.* (2011, 2013) use data from 1983 to 2009/2011 on investment projects and old structural adjustment and new policy-based loans. In addition to macroeconomic variables and basic project characteristics such as loan size, project length, sectoral dummies, and preparation and supervision costs, they deploy a sheaf of early warning indicators which are developed by the task team leader as the project evolves over its life (of approximately six to ten years). They find that loan size, project supervision costs, and certain early-warning indicators are significantly correlated with project outcomes. They also find that task team leader quality is significantly related to project success. Quality is measured by the average IEG rating of projects managed by the task team leader, other than the project in question. Denizer *et al.* (2012) find that investment projects in fragile and conflict-affected states show essentially the same characteristics as those in non-fragile states.

A difficulty with Kilby (1995, 2000), Wane (2004), Chauvet *et al.* (2010), and Denizer *et al.* (2011, 2012, and 2013), however, is that they do not adequately distinguish investment projects from policy-based loans. As explained earlier, policy-based loans support a set of policy and institutional reforms. They do not directly finance physical infrastructure and are not earmarked as are investment projects. Policy-based loans are shorter in time span and all prior actions/conditions are met before the presentation of the loan to the World Bank Board of Executive Directors. Due to their distinct nature, policy-based loans have a different concept and time-frame[4] as regards supervision and implementation, and involve different criteria for evaluation compared to investment projects. The only nod to the distinctive nature of policy-based lending in the latter studies is the inclusion of a dummy. Policy-disbursing loans are indeed a small fraction of the total loans of the World Bank in terms of numbers in most samples of the literature reviewed, e.g. 10 percent in Chauvet *et al*. (2010, p. 23) and in Denizer *et al*. (2013, p. 293). Hence these papers can be taken as representative of investment loans but they apply less to policy-based lending.

[4] Time-frames: A stand-alone development policy operation typically closes one year after its approval by the Board. A programmatic series typically closes one year after the approval of the last operation (loan) in the series. The Implementation Completion Review (ICR) is currently completed 12 months after closure, but the loans that are part of the dataset used for this paper had to complete their ICR 6 months after closure.

Additionally, Denizer *et al.* (2013) do not distinguish between the old style structural adjustment loans (pre-2004) and the newer development policy lending (post-2004), which have marked differences. Denizer *et al.* (2012) argue that "investment projects do better than policy lending operations" because they note a positive and significant coefficient on a variable for "investment projects" (the left-out category being fast-disbursing lending). Denizer *et al.* (2013) also find a positive coefficient on the "investment projects" dummy, although this is significant at the 10% level only in a data set for 1983-2011, whereas using a data set for 1995-2011 it is statistically insignificant, as they acknowledge (p. 293). There is a straightforward explanation for these findings: the failure rate of adjustment loans was higher, at 30.2% (July 1985 to June 2004)[5], than that of development policy lending at 19.6% on average (from September 2004 onwards)[6].

The present paper adopts this third type of approach that examines the impact of official financing on specific program/project development results/outcomes. One advantage of this approach is that the larger sample sizes (i.e., the number of projects/programs, often greatly exceeding the number of countries) permit more powerful statistical testing. Another advantage is that this approach enables a focus on quality of the aid/financing delivered.

This paper attempts to contribute further to the understanding of the impact of development financing on intended development outcomes at the program level by focusing solely on policy-based lending, and on these operations post-2004. It also attempts to provide further refinements in specifications to important variables used in the literature. For example, calculating a measure of the quality of a task team leader (TTL) is complicated by the fact that projects often have multiple TTLs over their life span, and the World Bank's electronic data system does not necessarily list the task team leader(s) that took a project or operation to the Board. As a result, previous studies have been limited to the track records of the task team leaders responsible for implementation, rather than identification and preparation, which are more relevant for the outcome of policy-based loans.

## 3. Data, variables and methodology

[5] Computations using the Bank's electronic record system. "Failure" is defined as in footnote 3.

[6] Using the data set constructed for the present paper.

Our research question is, “Are there design elements and other factors of policy-based loans which are tightly linked with success in achieving the intended development outcomes?” It should be emphasized that we do not seek to investigate the long-term or ultimate impact of development policy lending. Our focus is narrower. On the maintained assumption that the government, together with the World Bank’s teams, has identified the key areas for reform and the key results, we examine the factors that are associated with success in achieving these intended results.

### *3.1 Data*

The data sources are: (i) records maintained by the Operation Policy and Country Services unit of the World Bank and recently expanded for the purpose of this paper on all development policy lending (i.e., policy-based loans after the 2004 reform); (ii) the individual program documents of all development policy loans; (iii) the ICRs of each loan; (iv) the IEG validation of ICR ratings; (v) data extracted from the World Bank electronic record system; (vi) data on Country Policy and Institutional Assessment (CPIA) of the World Bank; and (vii) the World Bank database for the macroeconomic variables. There are 312 operations in the sample, spread across 92 countries and disbursing US$57.9 billion in total.

### *3.2 Dependent variable*

As in most of the related research, we use the World Bank’s Independent Evaluation Group (IEG) validation of the ICR rating as the outcome or dependent variable. IEG produces one rating for each loan, and has since 2010 produced one for each programmatic series. We construct a Likert scale from the IEG rating (highly unsatisfactory = 0, unsatisfactory = 1, moderately unsatisfactory = 2, moderately satisfactory = 3, satisfactory = 4, and highly satisfactory = 5) [7]. Of the entire 312-strong sample, 8 were highly satisfactory, 132 satisfactory, 111 moderately satisfactory, 42 moderately unsatisfactory, and 19 unsatisfactory.

The distribution of outcome ratings by region is presented in Figure 1 (Annex 1). Sub-Saharan Africa is by far the most active region, with 106 operations, followed by Latin America and the Caribbean with 71. The majority of ratings – 78 percent – fall in the range “moderately

[7] No operations were rated highly unsatisfactory. Hence there is a five-point scale.

satisfactory" to "satisfactory". The regions with the greatest rate of failure – viz. "unsatisfactory" to "moderately unsatisfactory" – are Middle East and Northern Africa (32 percent) and Sub-Saharan Africa (25 percent). The region with the highest rate of success – viz. "moderately satisfactory", "satisfactory" or "highly satisfactory" – is Eastern Europe and Central Asia, at 92 percent.

In addition, we run the regressions using the ratings of the Implementation and Completion Results Report (ICR). The latter cannot be considered independent of the Bank's management.

### *3.3 Variables of interest*

We hypothesize that the success of policy-based loans post-2004 depends heavily on key elements of reform program design. Below we present a short explanation of each; precise definitions are given in Annex 2. Illustrative examples drawn from actual operations are listed in the *Correlates of Success Codebook* which can be shared upon request.

*Weaker prior actions.* We hypothesize that prior actions or "conditions" agreed upon with the country authorities that are "substantive", viz. policy- and action-oriented, are more likely to attain their expected goals; and that prior actions that lack policy substance and are less action-oriented are less likely to attain their goals. For example, focusing on process-oriented steps, such as setting up "task forces", doing studies, and issuing "plans" for the future, may be less likely to attain tangible results, particularly if supported in later stages of programmatic series of loans without actionable policy as follow up. Based on desk reviews of Program Documents, the fraction of weaker prior actions is calculated for each operation.

*Matrix/results consistency*. The Bank requires development policy loans to be underpinned by a results framework, which provides the basis for monitoring and evaluation. A desk review of Program Documents was performed to construct a variable for each loan and programmatic series of loans that examines the extent to which there is consistency, or a clear "line of sight", between the prior actions or conditions for the program, the stated objectives of the program, and the intended results/outcome indicators. In the variable constructed, consistency occurs when the program objectives are all measured by meaningful results/outcome indicators, and where there is a clear line of sight from prior actions or conditions for the loan to the intended results/outcomes that are evaluated at the latter stage. If there are no result indicators to assess whether a particular program development objective has been attained, then the program may receive a lower rating. Alternatively, if the result indicators are of too high an order, it may be difficult to attribute them to the prior actions. For example, a prior action to improve education financing might have a result measure on reducing poverty from x to y, but within the timeframe of the evaluation[8] it would be difficult and possibly inappropriate to attribute any observed reduction in poverty to that reform, because there are many other contributory factors or because such outcomes need more time to materialize. Furthermore, results which are very early outputs or restatements of the prior actions may have less impact and may not serve for evaluation, and so are likely to lead to a negative rating.

*Programmatic vs. stand-alone operations*. A dummy variable is included to identify if a loan is a stand-alone operation or part of a programmatic series of loans. Programmatic series of loans have two advantages: the government has assured funding for a period of two to five years, there is clarity to follow through with an established reform agenda, and the program results agreed have a longer gestation period and hence may be more ambitious and may be more likely to be achieved. On the other hand, stand-alone operations are selected when there is political opportunity to reform or a higher degree of medium term uncertainty. Under these circumstances the result measures may be less ambitious. It is an empirical matter as to whether the increased time associated with programmatic series gives more flexibility and hence a greater probability of success, or whether the increased ambition associated with programmatic series involves greater risk and hence a lower probability of success.

*Task team leader skills*. Care was taken to identify the task team leader *as of the time the program went to the board*, by reference to the original Program Documents. We create two variables: (a) the number of policy-based loans the task team leader has taken to the board, excluding the current operation; and (b) a track record consisting of the sum of the IEG ratings of the operations the task team leader has taken to the board, prior to the current operation. When there were two or more task team leaders, they were considered co-equal, so their experience (a) and their track record figures (b) were averaged. See Annex 2 for the details.

[8] For more detail on time-frames see footnote 4.

*Number of prior actions by sector as a fraction of all prior actions; and dummy variables by sector affiliation of the task team leader.* The former set is needed because different sectors may be characterized by differing levels of risk or difficulty in achieving development results. The latter tries to gauge if either more experience doing policy-based loans or better quality control of certain sectoral departments within the Bank have an impact.

### *3.4 Control variables*

*The log of per capita GDP.* This would likely proxy for several country development conditions, such as the level of skill and education of government employees, strength of institutions to implement reforms, and the predictability of the political process, among others.

*The macroeconomic cluster and/or the governance component of the Country Policy and Institutional Assessment (CPIA).* The justification for the former is that weaker macroeconomic policies and institutions jeopardize macroeconomic outcomes, which in turn may have a serious impact on many development outcomes. The justification for the latter is that poor governance – corruption, patronage, weak fiduciary systems – may hamper development outcomes even if the reforms, on paper, are undertaken.

*Force majeure.* We codified this dummy variable to indicate if major natural disasters or coup d'états occurred between the approval of the operation and the evaluation of results. These shocks are not under the control of the government that negotiated the loan or the lending organization. Yet they may have a significant impact on short to medium term development outcomes/results that could be expected as a consequence of certain reforms.

*IBRD.*[9] The higher level of development of IBRD member countries may make for more successful loans. This may happen because governments may be more stable or have increased capacity to implement reforms, which has produced their higher incomes in the first place. On the other hand, IBRD loans are likely to have more ambitious reforms and more ambitious result

[9] The World Bank's lending falls into two parts: the International Bank for Reconstruction and Development (IBRD) and the International Development Association (IDA). The former lends to countries with per-capita income above a cut-off level at terms slightly more attractive than commercial. The latter lends to countries below that cut-off level at highly concessional terms, including through grants.

measures (or higher targets) than are IDA credits, and so it is an empirical matter whether IBRD loans would be more successful.

### *3.5 Model specification*

As a starting point, a descriptive univariate analysis was conducted for 49 potential regressors. For each variable, a t-test for the difference in means between "successful" and "failed" policy-based loans was performed (see Table 1 in Annex 1). The results of the univariate analysis were used as a selection criterion for identifying control variables, in addition to the core regressors related to program design (e.g., matrix/results consistency, weaker prior actions, and task team leader skills) and basic country characteristics. In a preliminary set of OLS regressions, the regional department that delivered the operations, the sector departments in which the task team leader was affiliated, and the fraction of conditions of different sector departments in the total conditions of each loan are included as regressors.

Based on these results, we establish a basic model, which is examined for a number of specifications, using two different econometric models. First, we use a simple OLS (see results in Annex 1, Table 2). The model can be specified as follows:

$$y_i = \beta_1{}' x_{1i} + \beta_2{}' x_{2i} + e_{\boldsymbol{i}},$$

where $y_i$ is the outcome rating on a 5-point scale (see section 3.2) for operation $i$; $x_{1i}$ is a vector of observed key design elements (see details in section 3.3), $x_{2i}$ is a vector of country characteristics (details in section 3.4), and $\beta_1$ and $\beta_2$ are vectors of coefficients. We use cluster standard errors (denoted by $e_i$), throughout, because both the independent variables and the dependent variables of the programmatic series of operations are likely to be correlated.

Second, we use an ordered probit model, which assumes that each succeeding level in the dependent variable is *superior to* the last, but without specifying a *numerical difference* as is implicit in the OLS model (see results in Annex 1, Table 3). The probability for a specific outcome rating based on a set of observable project characteristics and country characteristics can be expressed as:

$$Prob[y_i = 1 | x_{1i}, x_{2i}] = 1 - \Phi(\beta_1{}' x_{1i} + \beta_2{}' x_{2i} - u_1)$$

$$Prob[y_i = 2|x_{1i}, x_{2i}]$$
$$= \Phi(\beta_1{}'x_{1i} + \beta_2{}'x_{2i} - u_1) - \Phi(\beta_1{}'x_{1i} + \beta_2{}'x_{2i} - u_2)$$
$$Prob[y_i = 3|x_{1i}, x_{2i}]$$
$$= \Phi(\beta_1{}'x_{1i} + \beta_2{}'x_{2i} - u_2) - \Phi(\beta_1{}'x_{1i} + \beta_2{}'x_{2i} - u_3)$$
$$Prob[y_i = 4|x_{1i}, x_{2i}]$$
$$= \Phi(\beta_1{}'x_{1i} + \beta_2{}'x_{2i} - u_3) - \Phi(\beta_1{}'x_{1i} + \beta_2{}'x_{2i} - u_4)$$
$$Prob[y_i = 5|x_{1i}, x_{2i}] = \Phi(\beta_1{}'x_{1i} + \beta_2{}'x_{2i} - u_4)$$

where $\Phi(.)$ is the cumulative normal distribution function; and the threshold values ($u_1$, $u_2$, $u_3$, $u_4$), which are unknown, are estimated together along with the coefficients defined by the vectors $\beta_1$ and $\beta_2$. As in the earlier model, we compensate for cluster correlation between observations that are part of a programmatic series.

We apply robustness checks to the OLS models, in the form of least absolute deviations (LAD, or median) regression. The rationale is as follows: if the distribution of errors is Gaussian with contamination, then OLS, with its "breakdown point"[10] of 1/n, is highly sensitive to y-outliers and x-outliers. Robust estimators such as LAD are insensitive to y-outliers, so if the sign and magnitude of the OLS and LAD estimates are similar, we can at least be sure that we were not being misled by y-outliers. In respect of x-outliers, however, the breakdown point of LAD is likewise 1/n; but high-breakdown estimators such as least trimmed squares[11] are characterized by slow convergence and non-unique minima and there is as yet no computationally secure method, so the vulnerability to x-outliers cannot be satisfactorily addressed at the present time.

A further robustness test takes the form of adding a complete set of country dummies to the (otherwise) preferred estimate. This is a harsh way of eliminating country heterogeneity which might drive certain results, such as the coefficient on the skill variable (track record of the task team leader), which may be correlated with country-specific effects because assignments of task team leaders to a particular country often extend over several years.

[10] The breakdown point is the smallest fraction of contamination (e.g. by outliers) that can cause the estimator to take on values arbitrarily far from the true ones. For least squares, just one outlier is sufficient to move any coefficient an arbitrary distance away from the true, so that the breakdown point is 1/n, where n is the number of observations. See Rousseeuw and Leroy (1987, p. 9ff) for an exposition.

[11] The properties of Least Trimmed Squares are described by Rousseeuw and Leroy (1987), p. 15, and pp. 132-134.

## 4. Results

Table 1 in Annex 1 presents t-statistics on the difference in means for a large array of potential variables. In this analysis, "Successful" means a validated rating by IEG of "moderately satisfactory" or better; "failed" means a rating of "moderately unsatisfactory" or poorer. Table 2 in Annex 1 presents the OLS regression results. These are similar, in sign and significance, to the results of the ordered probit models in Table 3.

The matrix consistency variable, that is, having a clear line of sight between policy reforms and intended development results, enters all model specifications with a large positive and highly statistically significant effect on performance of development policy loans. Having a deep understanding of the likely impact of specific policy and institutional reforms seems critical for these operations. It also points to the need for realism in setting development results, considering the specifics of the reforms supported, country circumstances, and timeframe for the fruition of reforms into concrete results vis-à-vis the evaluation timetable. Realism, however, should not be interpreted as lack of ambition or risk aversion, but as having a clear understanding of the line of sight between reforms and results.

The coefficient on weaker prior actions is large and negative, that is, having a larger fraction of "weaker prior actions" in a loan hampers the performance of loans in achieving the intended outcomes. This result is statistically significant in some model specifications. However, when the variable on matrix consistency is taken out of the model, the statistical significance of weaker prior actions improves. This result hints at the importance of having reforms or policy measures that are actionable and that can indeed lead to tangible results as expected. There may be further reasons why weaker prior actions seem to increase the likelihood of failure. For example, an accumulation of such actions may be a signal of a government that is not really committed to serious reform.

The variables created to examine differential skills or capacity across different sectors/departments of the World Bank deliver some interesting results, though these need to be interpreted with caution. The dummy for the former Economic Policy sector (a former department within a larger unit in each region, now in substantial part, gathered under a single global practice called Macroeconomics and Fiscal Management) is positive and statistically significant in all model specifications. The economic policy sector brings together economists –

mostly but not exclusively macroeconomists – who often work, or have worked, as “country economists”, with a wide array of responsibilities including the dialog on macroeconomic, structural and growth policies with country authorities. They also undertake a sizable portion of the economic analytical work at the country level. On account of having an integrative role” vis-à-vis the policy reform agenda, economists in this sector have more often than not been put in charge of policy-based lending. Most Economic Policy sector staff have more experience with development policy operations than do other sectoral staff. For example, the average Economic Policy task team leader had taken 1.3 policy-based loans to the Board while that average is 0.4 for the environment department. This may have also fostered a more intensive “learning by doing” for teams and for managers as regards quality control.

The results for other sectors/departments of the World Bank in the t-tests on difference in means were negative and statistically significant for the energy, water, health, and public sector governance sectors. In the initial OLS regressions the negative and statistically significant result holds only for the energy sector. Overall, these results may be associated with limited experience, given that some sectors/departments have led relatively few of these operations, even though the substance of the actions/conditions within those sectors formed part of the reforms supported by loans led by other departments. To test this further we use the variable defined as the fraction of conditions in specific sectors. We find that the fraction of prior actions in the energy sector was consistently negative and strongly significant. This result should probably be interpreted in the context of the political difficulties, inherent complexity, time needed to generate results, or higher risk of reversal in this sector. Moreover, given the importance of fiscal and structural policies in the energy sector, they can be viewed as high risk-high return measures. At the same time, an in-depth loan-by-loan review showed that a significant portion of the prior actions in this sector tended to be weaker although the expected results were ambitious (i.e., hard to attain based on the measures supported).

The coefficients on the fraction of prior actions in loans for transport, agricultural, and water sectors were also negative and statistically significant in some specifications, but their significance eroded in the robustness checks.

We find that task team leader skills, as defined by the average rating of past operations, are a factor in the success of policy-based operations. This result holds in all model specifications.

Even after stripping out the effects of inter-country heterogeneity, the variable remains statistically significant although reduced in size. This is consistent with Denizer *et al.* (2013). On the other hand, the *number* of past operations taken to the board by the task team leader had limited to no impact on program outcomes. There is more to explore in this area, but initially it appears that the quality of the work done by the task team leader – as reflected in the rating of the recent operations – is the determining factor. The robustness of the task team leader track record needs to be examined with caution, however, on account of the possibility that task team leaders in particular countries may benefit or lose from the level commitment of the countries involved to reform implementation.

The variables on programmatic engagements and loan size did not show any statistical significance. Based on our qualitative review of program documents we observe that the more opportunistic stand-alone loans typically have shorter term results that tend to be less ambitious and easier to achieve than those of programmatic engagements (that have two loans or more). At the same time, reforms in programmatic engagements can be followed up from gestation to full implementation. Perhaps this result shows how these two factors counter-balance each other when performing regression analysis. Other variables tested, such as the regional dummies (accounting for the 6 regional departments of the World Bank) showed statistical significance only in the t-test on difference in means (see Table 1). However, the statistical significance did not hold in the OLS and ordered probit regressions, presumably as other variables absorbed their effect.

The lack of statistical significance of the CPIA variables— viz. the components for macroeconomic management and governance— may be surprising at first sight and not aligned with the findings of studies on the correlates of investment lending (e.g., Denizer *et al.* 2013). However, two important factors may explain this situation. First, unlike investment loans or the old style structural adjustment loans— which are used as data in the cited literature— the new policy-based loans (post 2004) require a country to have an adequate macroeconomic policy framework. This requirement excludes countries that do not fulfill that requirement at a particular time. Second, while these loans do not have a minimum threshold for "governance" for the recipient country, analysis undertaken under the World Bank's DPL Retrospective (2012) suggests that the larger share of policy-based loans between 2005 and 2011 went to countries with better governance and fiduciary ratings in the CPIA. Moreover, countries with stronger governance and fiduciary systems received a larger share of the total World Bank financing in the form of policy-based loans while countries with weaker governance (low CPIA ratings) received most of their financing in the form of investment projects.

The variable "*force majeure*" which was designed to capture major natural disasters and coup d'états, that occurred over the period between approval and the evaluation period, is statistically significant and has a negative coefficient, as expected. It suggests the vulnerability of some developing countries, even at the policy level, to these shocks.

The robustness of the results was checked using a quantile (50%) regression, which minimizes the sum of the absolute residuals rather than the sum of the squared residuals (table 4). The coefficient estimates are similar to those presented earlier, but the standard errors are somewhat larger, and some of the variables, notably the dummies for prior actions in agriculture and water, have their significance eroded. In this type of regression, country heterogeneity may also affect certain coefficients. This possibility is tested by using the harsh approach of applying a full set of country dummies, thereby eliminating all inter-country variation (see Table 4 in Annex 1).

Additionally, in all cases, two regressions were run, one with the IEG validated rating as the dependent variable and the other with the ICR original rating. Since the correlation between the two is high, very similar results emerged, and so the latter are not reported.

## 5. Conclusion

This paper examines factors that are correlated with the success of policy-based loans of the World Bank in achieving its intended development results. These loans represent roughly 30 percent of the institution's commitments and are a core platform for policy reform dialogue between the World Bank and country authorities. As highlighted by World Bank's internal reviews, this lending instrument has proved to be flexible and effective in supporting needed reforms and pursuing important development outcomes. Its success rate in achieving intended results is high (81.4% for the period September 2004 to July 2012). Thus, the findings of this paper highlight opportunities to keep up the good performance, particularly as reform complexity is likely to increase under the second, third and fourth waves of reforms in emerging and developing economies.

The findings have relevance for both the World Bank and for country policy makers engaged in policy dialogue. For example, intensified efforts could be made in the joint work of Bank teams and country authorities to improve the consistency, or "line of sight", between the reforms agreed and the results intended, striking the right balance between realism (including as regards to the time frame for achieving results) and ambition. Additional work to rigorously prune away low-quality results and prior actions with inconsistent results chains may be helpful. In relation to that, having fewer process-oriented prior actions (i.e., fewer "weaker prior actions"), particularly at later stages of the programmatic series of loans, may help to have more action-oriented reform packages that can deliver results.

Overall, these design improvements may also result in greater selectivity in the determination of prior actions/conditions, and possibly also in better selection of sectors within a country where policy reforms are likely to have strong impacts.

There is a growing consensus on the need to have well-trained and right-skilled team leaders at the helm of these operations. Further training, re-training, and other measures could help in fostering task team leader skills. More systematic pairing arrangements in teams with successful team leaders, as is done in some of the Bank's regional departments, may be helpful.

Sector affiliation of the team leader also seems to matter, as shown by the positive impact of having the operation led by the former Economic Policy sector. We have two assumptions in this regard. First, the experience of country economists as "dot connectors" of the reform agenda may be critical, particularly in the context of the cross-sectoral nature of reforms in developing and emerging economies. For example, a reform of energy subsidies  is linked not only to structural issues in the energy sector and the targeting of social assistance for the poor to cope with price adjustments, but also to the fiscal framework and expenditure efficiency decisions. A key role for the economists that integrate all these issues can be important for the success of such an operation. Second, the more intensive "learning by doing" of this department may be a factor as well, but to accelerate the process a more dynamic exchange of experiences could be fostered.

This may also hint at the need for training and retraining opportunities, as well as a more formal accreditation system for task-team leaders.

The fact that energy-heavy operations may increase the risk of having less successful loans needs to be interpreted carefully. As discussed earlier, these operations tend to be high-risk high-reward and by no means should be avoided. On the contrary, given the criticality of this sector for developing and emerging economies, efforts to move forward sector reforms need to continue. Deploying skilled teams and having thorough discussions in the review process may help in this endeavor.

## References

Boone, Peter, 1995. *Politics and the effectiveness of foreign aid.* Center for Economic Performance, London School of Economics and Political Science. Discussion Paper No. 272. December 1995.

Burnside, Craig, and David Dollar, 2000. "Aid, policies, and growth." *American Economic Review*, 90(4): 847-868.

Burnside, Craig, and David Dollar, 2004. *Aid, policies, and growth: revisiting the evidence.* World Bank Policy Research Working Paper 3251. March 2004. 36 pages.

Chauvet, Lisa, Paul Collier and Marguerite Duponchel, 2010. *What explains aid project success in post-conflict situations?* World Bank Policy Research Working Paper no. 5418, September 2010. 26 pages.

Collier, Paul, and David Dollar, 2002. "Aid allocation and poverty reduction". *European Economic Review* 46 (2002) 1475-1500.

Clemens, Michael A., Steven Radelet, Rikhil R. Bhavnani and Samule Bazzi, 2004. *Counting chickens when they hatch: timing and the effects of aid on growth.* Center for Global Development, Working Paper 44, July 2004.

Deininger, Klaus, Lyn Squire and Swati Basu, 1998. Does economic analysis improve the quality of foreign assistance? *The World Bank Economic Review*, 12(3): 385-418.

Denizer, Cevdet, Daniel Kaufmann, and Aart Kraay, 2011. *Good countries or good projects? Macro and micro correlates of World Bank project performance.* World Bank Policy Research Working Paper 5646, May 2011. Development Research Group, Macroeconomics and Growth Team.

Denizer, Cevdet, Daniel Kaufmann and Aart Kraay, 2012. *Good countries or good projects? Macro and micro correlates of World Bank project performance: Supplementary results on fragile and conflict-affected states and face time*. November 2012. 6 pages.

Denizer, Cevdet, Daniel Kaufmann and Aart Kraay, 2013. Good countries or good projects? Macro and micro correlates of World Bank project performance. *Journal of Development Economics* 105: 288-302.

Dollar, David and Jakob Svensson, 1998. *What explains the success or failure of structural adjustment programs?* Macroeconomics and growth group, The World Bank, April 1998. Also: *Economic Journal* 110(466): 894-917, October.

*Evaluation of general budget support: Synthesis report*, 2006. International Development Department (University of Birmingham) and Associates. May 2006.

Geli, Patricia, Aart Kraay and Hoveida Nobakht, 2013. Predicting active World Bank project outcomes. Processed. September 2013. 12 pages.

Gould, W., and W. H. Rogers, 1994. Quantile regression as an alternative to robust regression. *Proceedings of the Statistical Computing Section*. American Statistical Association.

Independent Evaluation Group (IEG), World Bank, 2008. *Public sector reform: what works and why? An IEG evaluation of World Bank support.* Washington, DC. 90 pages.

Independent Evaluation Group (IEG), World Bank, 2010. *Poverty reduction support credits: an evaluation of World Bank support.* Washington, DC. 123 pages.

Joint Evaluation of Budget Support, 1994-2004. *Evaluation of general budget support: synthesis report.* May 2006. Team leader Stephen Lister. International Development Department (IDD), Birmingham; Mokoro; and many others.

Kilby, Christopher, 1995. Supervision and performance: the case of World Bank projects. Center for Economic Research, Tilburg University, The Netherlands. March 1995. 94 pages.

Kilby, Christopher, 2000. Supervision and performance: the case of World Bank projects. *Journal of Development Economics* 62 (1), pp. 233-259.

Kilby, Christopher, 2013. The political economy of project preparation: an empirical analysis of World Bank projects. *Journal of Development Economics* 105: 211-225.

ODI, *Sector budget support in practice: synthesis report,* 2010. Mokoro and Overseas Development Institute. Authors Tim Williamson and Catherine Dom. London. February 2010.

Rousseeuw, Peter J., and Annick M. Leroy, 1987. *Robust regression and outlier detection.* New York: John Wiley and Sons. 329 pages.

Smets, Lodewijk and Stephen Knack, 2014. *World Bank lending and the quality of economic policy.* World Bank Policy Research Working Paper no. 6924, June 2014.

Wane, Waly, 2004. *The quality of foreign aid: country selectivity or donors' incentives?* World Bank Policy Research Working Paper 3325, June 2004. 33 pages.

World Bank, 2005. "Harmonized evaluation criteria for ICR and OED evaluations". 11 pages. October 6, 2005.

World Bank, 2011. Good practice note for development policy lending: Designing development policy operations. Operations Policy and Country Services, January 2011.

World Bank, 2013. *2012 Development Policy Lending retrospective: results, risks and reforms.* Operations Policy and Country Services.

## Annex 1. Descriptive statistics, figures and regression results

**Table 1. Features of successful and failed development policy operations, 2004-2012. Significance levels of t-tests are shown in the last column: *p<0.10, **p<0.05, ***p<0.01.**

| | *Successful* | *Failed* | *t-statistic on difference in means* |
|---|---|---|---|
| **Country characteristics** | | | |
| CPIA score (overall) | 3.7 | 3.6 | 0.75 |
| CPIA score (macroeconomic cluster) | 4.0 | 3.9 | 0.88 |
| CPIA score (governance cluster) | 3.4 | 3.4 | 0.54 |
| GDP per capita (real 2005 USD) | 2421 | 1835 | 1.63 |
| IBRD | 47.0% | 33.3% | 1.95* |
| *Force majeure* | 0.0 | 0.13 | -6.21*** |
| *Regional dummies* | | | |
| AFR | 31.0% | 43.3% | -1.83* |
| EAP | 9.9% | 6.7% | 0.78 |
| ECA | 18.7% | 6.7% | 2.27** |
| LCR | 21.8% | 18.3% | 0.59 |
| MNA | 7.1% | 15.0% | -1.95* |
| SAR | 11.5% | 10.0% | 0.33 |
| **Task team leader characteristics** | | | |
| Number of task team leaders | 1.37 | 1.23 | 1.62 |
| Number of previous operations taken to board | 1.12 | 1.03 | 0.49 |
| Track record (average IEG rating of previous operations) | 4.45 | 3.70 | 5.16*** |
| ***Sector/Department affiliation***: | | | |
| Economic Policy | 53.6% | 31.7% | 3.10*** |
| Poverty Reduction | 9.5% | 16.7% | -1.60 |
| Public Sector Governance | 9.1% | 16.7% | -1.71* |
| Financial and Private Sector Development | 9.1% | 5.0% | 1.04 |
| Social Development | 0.4% | 0.0% | 0.49 |
| Urban Development | 2.0% | 3.3% | -0.63 |
| Environment | 3.6% | 0.0% | 1.49 |
| Energy and Mining | 0.8% | 8.3% | -3.6*** |
| Education | 4.4% | 6.7% | -0.74 |
| Social Protection | 4.8% | 6.7% | -0.60 |
| Health, Nutrition and Population | 0.0% | 1.7% | -2.06** |
| Water | 0.0% | 1.7% | -2.06** |
| Agriculture and Rural Development | 1.6% | 0.0% | 0.98 |
| Transport | 0.8% | 1.7% | -0.62 |
| Financial Management | 0.4% | 0.0% | 0.49 |
| **Operation characteristics** | | | |
| *Structure:* | | | |
| Programmatic | 57.0% | 54.1% | 0.40 |
| Loan size (current US$ millions) | 198.2 | 133.5 | 1.60 |
| *Prior actions:* | | | |
| Number of prior actions | 10.5 | 10.4 | 0.16 |
| *Prior actions by sector as % of total in each loan* | | | |
| Agriculture, forestry and fisheries | 3.5% | 4.1% | -0.47 |
| Public administration, law and justice | 58.7% | 51.6% | 1.69* |
| Information and communications | 1.1% | 0.5% | 1.01 |
| Education | 6.8% | 8.4% | -0.68 |
| Finance | 10.2% | 5.4% | 1.66 |
| Health | 6.9% | 9.5% | -1.28 |
| Industry and trade | 4.1% | 2.4% | 1.46 |
| Energy and mining, of which: | 5.4% | 11.2% | -2.75*** |
| Energy | 4.2% | 10.0% | -3.16*** |
| Transportation | 2.0% | 3.8% | -1.34 |
| Water, sanitation and flood protection | 1.3% | 3.0% | -1.53 |
| Weaker prior actions, over total prior actions | 14.8% | 22.8% | -3.59*** |
| *Results framework:* | | | |
| Number of result measures | 24.4 | 22.3 | 0.71 |
| Vaguely stated result measures, as % of total | 8.6% | 8.5% | 0.07 |
| Results lacking baselines or targets, as a % of all results | 32.6% | 30.6% | 0.47 |
| Matrix/results consistency | 75.5% | 66.2% | 3.57*** |
| All operations (number) | 252 | 60 | - |

**Table 2: OLS Regression results**

| Model specification # | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Matrix/results consistency | 1.31***<br>(4.5) | 1.31***<br>(4.5) | 1.30***<br>(4.5) | 1.32***<br>(4.5) | 1.37***<br>(5.0) | |
| Weaker prior actions / total | -0.54<br>(-1.6) | -0.54<br>(-1.6) | -0.53*<br>(-1.6) | -0.62*<br>(-1.8) | | -0.77**<br>(-2.4) |
| Economic Policy Sector | 0.26**<br>(2.5) | 0.26**<br>(2.5) | 0.26**<br>(2.5) | 0.26**<br>(2.6) | 0.29***<br>(2.8) | 0.22**<br>(2.1) |
| Prior actions in agriculture / total | -1.34**<br>(-2.3) | -1.34**<br>(-2.3) | -1.31**<br>(-2.3) | -1.36**<br>(-2.4) | -1.44**<br>(-2.5) | -1.28**<br>(-2.0) |
| Prior actions in energy / total | -1.1***<br>(-3.1) | -1.1***<br>(-3.0) | -1.1***<br>(-3.1) | -1.1***<br>(-2.8) | -1.19***<br>(-3.3) | -1.2***<br>(-3.4) |
| Prior actions in transport / total | -0.66*<br>(-1.9) | -0.66*<br>(-1.9) | -0.64*<br>(-1.8) | -0.66*<br>(-1.8) | -0.74**<br>(-2.0) | -0.76**<br>(-2.1) |
| Prior actions in water / total | -0.46<br>(-1.6) | -0.46<br>(-1.6) | -0.46<br>(-1.6) | -0.51*<br>(-1.7) | -0.45<br>(-1.5) | -0.64*<br>(-1.95) |
| TTL track record | 0.44***<br>(4.2) | 0.44***<br>(4.2) | 0.44***<br>(4.4) | 0.44***<br>(4.3) | 0.46***<br>(4.6) | 0.45***<br>(4.1) |
| Programmatic | | -0.0029<br>(0.03) | | | | |
| Log loan size | | | 0.025<br>(0.7) | | | |
| CPIA Cluster Macro | 0.031<br>(0.4) | -0.031<br>(1.0) | 0.016<br>(0.2) | | 0.029<br>(0.3) | 0.059<br>(0.7) |
| CPIA Cluster Governance | | | | 0.177<br>(1.2) | | |
| Log per capita GDP | 0.019<br>(0.4) | 0.019<br>(0.4) | 0.011<br>(0.2) | -0.027<br>(-0.4) | 0.024<br>(0.4) | 0.057<br>(1.0) |
| *Force majeure* | -1.8***<br>(-10.0) | -1.8***<br>(-10.1) | -1.8***<br>(-9.7) | -1.8***<br>(-9.6) | -1.8***<br>(-9.5) | -1.7***<br>(-12.0) |
| Constant | 0.235<br>(0.4) | -0.235<br>(0.4) | 0.228<br>(0.7) | 0.094<br>(0.2) | -0.0091<br>(-0.02) | 0.84<br>(0.2) |
| Adj. R-sq. | 0.39 | 0.39 | 0.39 | 0.39 | 0.38 | 0.32 |
| N | 312 | 312 | 312 | 312 | 312 | 312 |

Method: OLS. Dependent variable: Likert scale of the IEG rating validation of the development policy operation (HS=5, S=4, MS=3, MU=2, U=1). * Significance at 10%, **significance at 5%, *** significance at 1%. Numbers in parentheses are cluster t-statistics, adjusting for correlation among operations within programmatic series.

## Table 3: Ordered Probit Results

| Model specification # | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | | | | | | |
| Matrix/results consistency | 1.97*** | 1.97*** | 1.96*** | 2.02*** | 2.06*** | |
| | (4.9) | (4.8) | (4.9) | (4.8) | (5.3) | |
| Weaker prior actions / total | -0.84* | -0.84* | -0.83* | -1.03** | | -1.13*** |
| | (-1.8) | (-1.8) | (-1.8) | (-2.2) | | (-2.7) |
| Econ. Policy Sector | 0.38** | 0.38** | 0.39** | 0.41** | 0.44*** | 0.30* |
| | (2.3) | (2.3) | (2.3) | (2.4) | (2.7) | (1.9) |
| Prior actions in agriculture / total | -1.92*** | -1.93*** | -1.90*** | -1.98*** | -2.08*** | -1.73** |
| | (-2.7) | (-2.7) | (-2.7) | (-2.8) | (-2.9) | (-2.3) |
| Prior actions in energy / total | -1.56*** | -1.55*** | -1.57*** | -1.44*** | -1.67*** | -1.64*** |
| | (-3.2) | (-3.1) | (-3.2) | (-2.9) | (-3.5) | (-3.4) |
| Prior action transport / total | -1.57* | -1.56* | -1.53* | -1.67* | -1.70** | -1.48* |
| | (-1.9) | (-1.9) | (-1.9) | (-1.9) | (-2.1) | (-1.9) |
| Prior actions in water / total | -0.74* | -0.74* | -0.74* | -0.84** | -0.71* | -0.95** |
| | (-1.9) | (-1.9) | (-1.9) | (-2.0) | (-1.9) | (-2.4) |
| TTL track record | 0.68*** | 0.68*** | 0.69*** | 0.69*** | 0.70*** | 0.66*** |
| | (4.4) | (4.4) | (4.5) | (4.4) | (4.7) | (4.2) |
| Programmatic | | 0.015 | | | | |
| | | (0.09) | | | | |
| Log loan size | | | 0.028 | | | |
| | | | (0.47) | | | |
| CPIA Cluster Macro | 0.061 | 0.057 | 0.045 | | 0.058 | 0.089 |
| | (0.51) | (0.46) | (0.37) | | (0.47) | (0.73) |
| CPIA Cluster Governance | | | | 0.39 | | |
| | | | | (1.57) | | |
| Log per capita GDP | 0.061 | 0.062 | 0.051 | -0.037 | 0.068 | 0.111 |
| | (0.75) | (0.74) | (0.58) | (-0.37) | (0.82) | (1.38) |
| *Force majeure* | -2.61*** | -2.61*** | -2.56*** | -2.61*** | -2.58*** | -2.39*** |
| | (-7.2) | (-7.1) | (-6.9) | (-7.0) | (-7.0) | (-7.8) |
| Constant | - | - | - | - | - | - |
| | | | | | | |
| Adj. R-sq. [1/] | 0.19 | 0.19 | 0.19 | 0.19 | 0.18 | 0.15 |
| N | 312 | 312 | 312 | 312 | 312 | 312 |

Method: Ordered Probit. Dependent variable: Likert scale of the IEG rating validation of the development policy operation (HS=5, S=4, MS=3, MU=2, U=1). * Significance at 10%, **significance at 5%, *** significance at 1% . Numbers in parentheses are cluster t-statistics, adjusting for correlation among operations within programmatic series. [1/] Pseudo R-squared.

## Table 4: Robustness checks

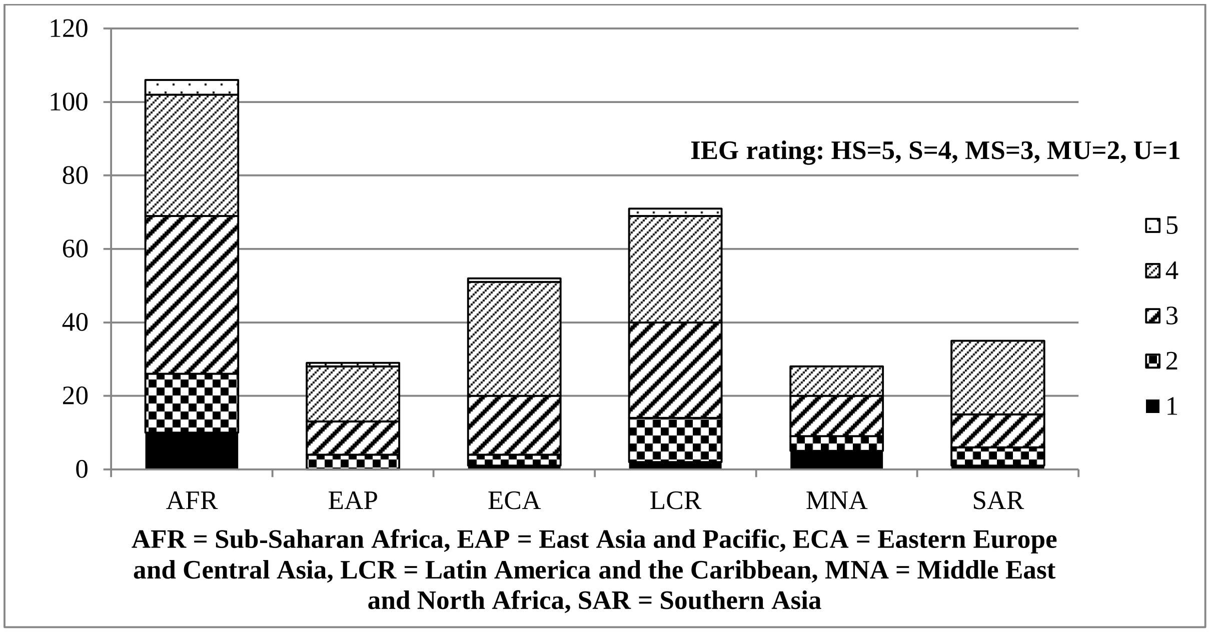
| Robustness specification | **(a)** | **(b)** |
|---|---|---|
| *Estimation method* | Median regression | OLS with country dummies [1/] |
| *Cluster standard errors?* | No | Yes |
| Matrix/results consistency | 1.04***<br>(3.3) | 0.79***<br>(2.8) |
| Weaker prior actions / total | -0.54<br>(-1.4) | -0.44<br>(-1.4) |
| Econ. Policy Sector | 0.30***<br>(2.8) | 0.28**<br>(2.0) |
| Prior actions in agriculture / total | -0.76<br>(-1.1) | -0.51<br>(-1.0) |
| Prior actions in energy / total | -1.37***<br>(-3.2) | -1.27***<br>(-3.0) |
| Prior actions in transport / total | -0.32<br>(0.6) | -0.83**<br>(-2.0) |
| Prior actions in water / total | -0.56<br>(-0.8) | -0.59<br>(-1.4) |
| TTL track record | 0.65***<br>(6.9) | 0.17**<br>(2.0) |
| Log gdp pc | 0.053<br>(1.0) | -1.36*<br>(-1.9) |
| CPIA Cluster A | 0.066<br>(0.7) | 0.15<br>(0.7) |
| *Force majeure* | -1.80***<br>(-5.4) | -1.16***<br>(-5.9) |
| Constant | -0.56<br>(-0.9) | 11.2**<br>(2.1) |
| Adj. R-sq. | 0.24 2[2/] | 0.57 |
| N | 312 | 312 |

Dependent variable: Likert scale of the IEG rating validation of the development policy operation (HS=5, S=4, MS=3, MU=2, U=1). * Significance at 10%, **significance at 5%, *** significance at 1% . Numbers in parentheses are t-statistics.

[1/] 92 country dummies absorbed.

[2/] Pseudo R-squared.

**Figure 1. Number of development policy operations, September 2004 to 2012, by IEG outcome rating and by region**



**AFR = Sub-Saharan Africa, EAP = East Asia and Pacific, ECA = Eastern Europe and Central Asia, LCR = Latin America and the Caribbean, MNA = Middle East and North Africa, SAR = Southern Asia**

## Annex 2. Definitions of variables and sources of data

For ease of reference, the variables are listed alphabetically.

Country Policy and Institutional Assessment (CPIA). The CPIA is intended to capture the quality of a country's policies and institutional arrangements. Two different sets of scores are used: a cluster average for Economic Management (CPIA questions 1-3), and a cluster average for Public Sector Management and Institutions (CPIA questions 12-16). CPIA scores for calendar year X are used to explain the outcome of operations presented to the board in calendar year X+1. Thus, for instance, if Board presentation occurred in 2006, then the CPIA figure refers to the situation in 2005. The information in the CPIA thus reflected part of the information set available to the government, the World Bank team, and the board, at the time of decision.

GDP per capita[12]. Stated in constant 2005 US dollars, using the nominal 2005 exchange rate. It refers to GDP per capita in the year in which the development policy loan in question went to the board.

IBRD dummy. The motivation for using this variable is presented in the main text. The variable =1 if IBRD financing for the loan in question exceeded that from IDA and other sources, and =0 otherwise.

IEGoutcomeLikert. This is one of the main dependent variables. It is a numerical transformation of the ratings by IEG of the overall operation outcome: unsatisfactory = 1, moderately unsatisfactory = 2, moderately unsatisfactory = 3, satisfactory = 4, and highly satisfactory = 5.

The fraction of weaker prior actions. This variable is broadly described in the main text. These are defined as actions that involve processes or procedures or documents but which do not themselves generate results because they are purely inputs into other processes. These include (i) plans, statements of policy and statements of intention, (ii) studies, reports, and evaluations,

[12] World Bank Open Data: http://data.worldbank.org/

(iii) terms of reference, (iv) the creation of task forces of a temporary nature, (v) exact repetitions of prior actions from the previous operation. On the other hand, (v) ministerial decisions are considered substantive if they carry the force of law, thereby giving some assurance that the intended actions will be taken, but are considered weaker if they do not carry the force of law, because there is then less assurance of the outcome. Furthermore, (vi) certain frameworks, mechanisms and systems go beyond statements of intention because they are in effect *decision rules* which influence incentives and behavior immediately. For example, a new electricity tariff framework or a new petroleum pricing framework may have immediate impacts in the market. Decision rules of this kind are considered to be substantive. Detailed illustrative examples drawn from actual operations are presented in the separate document *Correlates of Success Codebook* which can be made available upon request.

Loan size. This is the total size of the loan, in millions of current US dollars, including both the IBRD and the IDA amount.

Matrix/results consistency. This variable is broadly described in the main text. Further detail is provided here. Matrix/results consistency is the fraction of results that entail consistency among program development objectives, prior actions and results measures. Matrix/results consistency involves three major requirements. (a) All the program development objectives are evaluated by the results indicators. (b) All the prior actions contribute to one or more of the results indicators. The link between the prior action and the result measure must be causal; it is not enough that the prior action be in the same functional area or subsector as the result measure. (c) The result measures must satisfy four standards of quality. (i) The result measure must be an outcome or an output, that is, it must not be a further input into the processing chain. Thus to gather some data or to set up a committee is not a result measure because these are inputs into subsequent processes which might deliver results. (ii) The result measure must not be a simple restatement, even if in different words, of the prior action; it must be different from, and must go beyond, the prior action. (iii) The result measure must be stated precisely. It can be a Boolean (Yes/No, True/False), or a quantifiable measure. The aim here is to ensure that the result measure can later serve as the basis for evaluation. If the result measure is vaguely stated so that it can be multiply interpreted at the evaluation stage, then it is not a reliable basis for evaluation. (iv) The output or outcome measure must not be at so high a level that attribution becomes dubious.

Measures of country-wide poverty are not good as outcome measures because they are caused by many different factors; similarly measures of GDP growth and private sector credit. For the purposes of this measure, the public finance management area is omitted, because it is often difficult to specify meaningful results that are significantly different from prior actions. For instance, there is broad agreement that bringing the internal audits up to date is a valuable prior action, and often a difficult one, but it would be very complex and needlessly costly to invent an outcome indicator reflecting the improved financial discipline and honesty of officials two or three years out. Detailed illustrative examples drawn from actual operations are presented in the separate document *Correlates of Success Codebook* which can be made available upon request.

Number of result indicators. This is the number of result measures listed in the results matrix in the program document.

Prior actions in energy. This is described in the main text.

Programmatic. This is =1 if the operation in question is part of a programmatic series of operations. It is =0 for stand-alone operations.

Regional dummies. These follow the Bank’s regions: AFR, EAP, ECA, LCR, and SAR.

Sector Board (or Network) dummies. This is the network affiliation of the task team leader. The networks, until 2014, were Financial and Private Sector Development (FPD), Human Development Network (HDN), Poverty Reduction and Economic Management (PREM) or the Economic Policy Sector Board, and the Sustainable Development Network (SDN) which in turn consisted of several Boards including the Energy and Mining Sector Board.

Task team leader experience. This is described in the main text. It is the number of operations that the task team leader has taken to the board prior to the operation in question. When there is more than one task team leader, a simple average of all task team leaders’ experience is used.

Task team leader track record. This is described in the main text. It is the sum of IEG ratings of the operations that the task team leader has taken to the board prior to the operation in question divided by the task team leader experience.

When there is more than one task team leader, all contributors' track records are averaged to provide a project-level track record with variable name IEGprojaver. Since no-one can have a track record upon taking one's first operation to the board, the variable is not defined in these instances. IEGprojaver will be computed if at least one task team leader has some prior experience.

For a separate specification of the regression analysis we replace the missing task team leader track records with averages of non-missing observations in the historic portfolio. Using this variable, ttlav, we compute two alternative variables: A task team leader track record ttltrackind_withAVER for operations when there is only one task team leader and, again, a project-level track record representing the average across collaborators (variable name: IEGprojaver_withAVER).