Predicting World Bank Project Outcome Ratings

Patricia Geli

Aart Kraay

Hoveida Nobakht

WORLD BANK GROUP

Development Research Group

Development Economics Vice Presidency

&

Operations Policy and Quality Department

Operations Policy and Country Services Vice Presidency

August 2014

Abstract

A number of recent studies have empirically documented links between characteristics of World Bank projects and their ultimate outcomes as evaluated by the World Bank's Independent Evaluation Group. This paper explores the in-sample and out-of-sample predictive performance of empirical models relating project outcomes to project characteristics

observed early in the life of a project. Such models perform better than self-assessments of project performance provided by World Bank staff during the implementation of the project. These findings are applied to the problem of predicting eventual Independent Evaluation Group ratings for currently active projects in the World Bank's portfolio.

This paper is a product of the Development Research Group, Development Economics Vice Presidency; and the Operations Policy and Quality Department, Operations Policy and Country Services Vice Presidency. 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 patriciageli@worldbank.org, akraay@worldbank.org, and hnobakht@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.

Predicting World Bank Project Outcome Ratings

Patricia Geli, Aart Kraay, Hoveida Nobakht
The World Bank

1818 H Street NW, Washington DC 20433, patriciageli@worldbank.org, akraay@worldbank.org, hnobakht@worldbank.org. Patricia Geli and Hovieda Nobakht are in the Operations Policy and Country Services Vice Presidency of the World Bank, and Aart Kraay is in the Development Research Group of the Development Economics Vice Presidency of the World Bank. We are grateful to Kyle Peters for his encouragement of this project and for helpful feedback, and to Luis Servén for comments. The views expressed here are the authors', and do not reflect those of the World Bank, its Executive Directors, or the countries they represent.

1. Introduction

A number of recent studies have empirically documented links between characteristics of World Bank projects and their ultimate outcomes as evaluated by the World Bank's Independent Evaluation Group (IEG) [1]-[9]. However, less is known about the predictive power of these estimated relationships. In this paper, we explore the in-sample and out-of-sample predictive performance of empirical models relating project outcomes to project characteristics observed early in the life of a project. We show that such models perform better than self-assessments of project performance provided by World Bank staff during the implementation of the project. We apply these findings to the problem of predicting eventual IEG ratings for currently active projects in the Bank's portfolio.

Development projects financed by the World Bank are subject to a comprehensive monitoring process over the course of project implementation. Staff self-assessments of progress toward achieving the project's development objective (DO) are a key ingredient in this process. These assessments are captured in the Implementation Status and Results (ISR) reports prepared regularly during project implementation, and include a rating of performance relative to the project's development objective (the ISR-DO rating). There is growing evidence that these interim assessments tend to be optimistic, in the sense that a substantial number of projects rated as successful in the ISR-DO ratings nevertheless go on to be rated upon completion as unsuccessful by IEG. [10]

This is a concern because it suggests that opportunities to take mid-course corrective actions on projects in difficulty are missed due to overly optimistic ISR-DO ratings. Independent assessments of project performance by IEG have been downgraded on average by 15 percentage points compared with the Bank's final self-assessment just before implementation completion [10]. A degree of optimism in self-assessments is also visible in the active portfolio: out of 1591 active Investment Project Financing (IPF) operations as of end-FY13, only 13 percent of projects are rated as unsatisfactory in the ISR-DO ratings, even though the historical average rate of unsatisfactory projects as rated by IEG over the period 1995-2012 is 27 percent. This paper builds on, and adds to, the previous literature by showing how a successful prediction model can help to reduce these "disconnect" rates, i.e. the proportion of projects evaluated by IEG to be unsatisfactory even though they were rated as satisfactory by the Bank at the time of exit. More generally, it can help to flag problems in currently active projects that might lead to unsuccessful outcomes, so that remedial steps can be taken in advance.

We first use historical data on 2729 IPF operations evaluated by IEG between 1995 and 2012 to estimate the relationship between a small set of project characteristics and project outcomes as assessed by IEG. These project characteristics include project size, preparation time, effectiveness delays, planned project length, country performance as measured by the CPIA ratings, and the track record of the project manager (known as a “task team leader” or TTL), as measured by IEG ratings on other projects managed by the same TTL. Most of the predictive power of the model for project outcomes comes from these last two variables.

We document the in-sample predictive power of this model by comparing its predictions with the actual IEG evaluations, and find that it can successfully predict between 40 and 46 percent of unsatisfactory project outcomes. While this is far from perfect, the prediction model does a substantially better job of predicting outcomes than do ISR-DO ratings. For example, ISR-DO ratings in the first quarter of the life of a project correctly anticipate only 3 percent of unsatisfactory IEG ratings, while those in the second quarter do so correctly only 17 percent of the time.

Given this evidence on the ability of the prediction model to anticipate unsuccessful project outcomes, we next use it to predict project outcomes in the portfolio of 1591 active IPF operations as of end-FY13. The predictions of the model are based on real-time-observable information on project characteristics as of end-FY13. We calibrate the prediction model to generate a 27 percent rate of predicted unsatisfactory projects, consistent with the historical experience between 1995 and 2012. This identifies a set 430 projects at risk of ultimately having unsatisfactory IEG ratings. Of these, only 77 are rated by the ISR-DO rating as making unsatisfactory progress towards meeting their development objectives. We document how the incidence of these predicted-to-be-unsatisfactory projects varies across regions and across projects in different stages of implementation.

2. Institutional Setting

This paper focuses on the Bank’s Investment Project Financing (IPF) loans, credits and grants, which can be provided for a wide range of activities aimed at creating the physical and social infrastructure necessary to reduce poverty and promote sustainable development. While each Bank-financed project is a country-led effort, Bank staff typically monitor progress and provide inputs at each stage of the project (illustrated by Figure 1). The preparation phase usually begins once a proposal to address a development issue is identified and discussed between the borrower and the Bank. The Project Concept Note (PCN) is followed by Project Appraisal Document (PAD), which is subject to

approval by the Bank's Board of Executive Directors. Once approved and made effective, a project is implemented over several years, with the project spending financed by disbursements on loans or grants provided by the World Bank. The time elapsed between approval and effectiveness is often used as a proxy for country ownership. This is because delays in project implementation frequently reflect difficulties in ensuring required participation from government counterparts in the borrowing country, and so longer-than-expected elapsed time between approval and effectiveness can proxy for weak borrowing country commitment to the project [9]. Once the project is complete, an Implementation Completion Report (ICR) summarizes World Bank staff views on the implementation of the project.

A key ingredient in the PAD is a statement of the Development Objective (DO) of the project. It outlines the specific expected outcomes of the project, and subsequently the success of the project is rated relative to its development objectives. The first official Bank document to rate a project's progress toward achieving its development objectives is the Implementation and Status Results (ISR) report, completed by the project team, led by a task team leader, and approved by the relevant Sector Manager. The ISR is a tool that provides information for project and portfolio monitoring, and task team leaders are required to prepare it at least twice a year to report on the status of the projects for which they are responsible, over the lifespan of each project. Each ISR identifies key issues and actions for management attention and rates the likelihood that the project will meet its development objectives. Within six months of project completion, the Bank produces an ICR, which assesses the extent to which the project has achieved its stated objectives.

Following the completion of the ICR, the Bank's Independent Evaluation Group (IEG) completes an independent review of the ICR using the evidence contained in the ICR. This ICR Review (ICRR) assigns its own rating, which can differ from the Bank's self-assessed ratings, and provides additional analysis. A subset of roughly 25 percent of completed projects is subjected to a subsequent more detailed Project Performance Assessment Report (PPAR) produced by IEG. These are substantially more in-depth analyses by IEG staff, based on additional research and data gathered by IEG, and often involving field visits to project sights to gain further insights as to the project's performance. These are typically completed one to three years following project completion. If a PPAR rating is available for a project, we take this as the ultimate outcome rating. If not, we use the rating provided in the ICRR.

Figure 1: Timeline of key project ratings and variables constructed for the prediction model. PCN=Project Concept Note, PAD=Project Appraisal Document, ISR=Implementation Status and Results Report, ICR=Implementation Completion Report and PPAR=Project Performance Assessment Report. Timeline is not to scale.

At all four assessment stages (ISR, ICR, ICRR, and PPAR), project success relative to the development objective is scored on a 6-point scale (Highly Unsatisfactory/Unsatisfactory/Moderately Unsatisfactory/Moderately Satisfactory/Satisfactory/Highly Satisfactory). In the empirical work that follows we will focus on a binary classification of projects into those that score Moderately Satisfactory or higher, and those that do not.

3. The Prediction Model

The prediction model estimates the probability that a project is rated satisfactory, based on a set of observable project characteristics, i.e.

$$(1) \quad Prob[y_i = 1|x_i] = \Phi(\beta'x_i)$$

where $y_i = 1$ if project i is rated moderately satisfactory or better; x_i is a vector of observed project characteristics; and $\Phi(\cdot)$ is the cumulative normal distribution function. In order to be useful as a prediction model, the explanatory variables included in x_i must be (a) available at the time that interim assessments of active projects are made, and (b) significantly correlated with project outcomes in the historical data. In addition, a key assumption is that the model linking project outcomes to project characteristics is stable over time.

Predictions of project performance are based on the fitted values from the probit model. Specifically, it is natural to predict that a project will be successful if the predicted probability $\Phi(\hat{\beta}'x_i)$ is greater than some threshold, or equivalently, if $\hat{\beta}'x_i > q$, where $\hat{\beta}$ are the estimated coefficients on the variables included in the probit regression. The choice of q determines how successful the model is at correctly predicting successful and unsuccessful project outcomes, as summarized in the following table.

Table 1: Interpretation of Predicted versus Actual Outcomes.

	Predicted Outcome	
Actual Outcome	Unsatisfactory ($\hat{\beta}'x_i < q$)	Satisfactory ($\hat{\beta}'x_i > q$)
Unsatisfactory ($y_i = 0$)	<i>Correctly Predicted Unsatisfactory</i>	<i>Type 1 Error</i>
Satisfactory ($y_i = 1$)	<i>Type 2 Error</i>	<i>Correctly Predicted Satisfactory</i>

The prediction model is useful if it correctly predicts both satisfactory and unsatisfactory outcomes, or equivalently, if it has a low rate of Type 1 and Type 2 errors. The choice of the threshold q determines how well the model does along these two dimensions. For example, if we set q to be a very large number, we will correctly predict most unsatisfactory projects to be unsatisfactory, but at the same time, we will make many Type 2 errors, i.e. we will also wrongly predict many satisfactory projects to be unsatisfactory. Conversely, if we set q very low, we will correctly predict most satisfactory project outcomes, but we will also wrongly predict many unsatisfactory projects to be satisfactory.

The choice of the threshold q depends on the tolerance for these two types of errors. Since this is difficult to specify in advance, we begin with a natural benchmark: we choose q to target a plausible overall predicted success rate for projects. For example, if in the historical data 73 percent of projects were rated satisfactory, then it is natural to set q as the 73th percentile of $\hat{\beta}'x_i$, so that the model predicts the same overall success rate as in the historical data.¹

¹ A standard rule is to select the threshold q so that the rate of Type 1 and Type 2 errors is the same. Applying this rule in our sample delivers a threshold that is nearly identical to the historical average success rate of projects.

4. Estimating the Prediction Model

We begin by estimating the prediction model using all IPF operations evaluated between 1995 and 2005.² This will allow us to assess the out-of-sample predictive power of the model by how well it predicts actual project outcomes between 2006 and 2012. The basic prediction model includes the following correlates of project outcomes: project size (as measured by the logarithm of the original commitment in dollars), preparation time (months from PCN review to project approval date), elapsed time between approval and effectiveness (measured in months), initially planned project length (months from approval date to initial estimated closing date), TTL track record (average IEG rating on other projects managed by the same TTL)³, and country policy performance (average country-level CPIA score during life of project). The sample consists of 1561 IPF operations for which these variables are all available.

The main significant correlates of project outcomes are country policy performance, and TTL track record, which both enter positively and highly significantly. This is consistent with the findings of [6] which documents a strong relationship between project outcomes and both of these variables in a larger set of World Bank projects. In addition, projects that are initially planned to take a long time are significantly less likely to be rated satisfactory. Preparation time and time elapsed between approval and effectiveness are negatively correlated with project outcomes, but not significantly so. Unlike in some of our earlier work, project size is not significantly correlated with outcomes. In the prediction

² We exclude World Bank projects providing general budget support rather than specific project financing from our analysis as the lifecycle of these projects is very different from the bulk of the World Bank's IPF operations.

³ Calculating this measure of TTL track record is complicated by the fact that projects often have multiple TTLs over their lifespan. We work with project-level data reflecting the identity of the TTL as of the end of each fiscal year for the life of each project in our sample. We then construct the TTL record variable as a weighted average of the IEG ratings of all of the projects the TTL managed, with weights proportional to the number of fiscal years for which the TTL was responsible for each project. For example, if a TTL managed one project for 4 years and it was rated satisfactory by IEG, and another project for just one year of its life and this project was rated unsatisfactory by IEG, then the TTL record variable assigns a weight of 80 percent to the satisfactory outcome and 20 percent to the unsatisfactory outcome. To avoid a mechanical correlation between TTL track record and project outcomes, we drop the project in question from this calculation in the historical portfolio data. This means that the TTL record variable is available only for projects managed by TTLs who show up as TTLs for at least two projects in the dataset. Projects managed by "first-time" TTLs are therefore dropped from the analysis. In the prediction exercise in the active portfolio, we base the TTL record variable on all available projects in the historical data, even if there is just one past evaluated project for a given TTL. Finally it is important to note that the TTL track record variable is based only on the identity of the TTL during the implementation period of the project, and not from the project identification and preparation stages. This is because we draw TTL data from the ISR sequences for projects, and these begin only during the implementation phase.

exercises that follow, we include all these variables in the prediction model even though not all of them are statistically significant at the 10 level or higher.

Table 2: Estimated Prediction Model

Probit regression	Number of obs	=	1561
	LR chi2(6)	=	153.84
	Prob > chi2	=	0.0000
Log likelihood = -825.00538	Pseudo R2	=	0.0853

iegsat	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lntotalcom	.015695	.0320012	0.49	0.624	-.0470261	.0784162
preptime	-.0029434	.0021185	-1.39	0.165	-.0070957	.0012088
effdelay	-.0101739	.0074377	-1.37	0.171	-.0247516	.0044038
planlength	-.0077196	.0021558	-3.58	0.000	-.0119449	-.0034943
tntlqual	1.032869	.1265106	8.16	0.000	.7849125	1.280825
cpiaav	.4187925	.0755984	5.54	0.000	.2706224	.5669625
_cons	-.8541903	.292965	-2.92	0.004	-1.428391	-.2799894

5. Predicting Project Outcomes

Table 3 summarizes the in-sample (first column) and out-of-sample (second column) predictive performance of the model. Specifically, the first column summarizes how well the fitted values from the probit regression estimated over the period 1995-2005 predict project outcomes during the same period, while the second column asks how well the fitted values predict project outcomes during the next six years, 2006-2012. For reference, the first row of the table provides the historical average rates of unsatisfactory projects. The subsequent rows present predictions from the benchmark model described above, followed by predictions using the model with different combinations of the most significant correlates, the CPIA ratings and TTL track record.

Looking first within the sample, 26 percent of projects (413/1561) were rated as unsatisfactory during 1995-2005. The full model, including all correlates, correctly predicts 46 percent of these unsatisfactory outcomes. The out-of-sample predictions are naturally somewhat less accurate, but still correctly predict 40 percent of unsuccessful projects. A model with TTL record as the only variable correctly predicts 42 (38) percent in sample (out of sample), whereas a model with the CPIA alone correctly predicts 36 percent. The predictive performance of the model that combines the CPIA and TTL

is comparable to the full model, indicating that these two variables account for most of the predictive power of the model.

Table 3: Predictive Power of the Prediction Model

	In-Sample	Out-of-Sample
Unsatisfactory by IEG:	(1561 projects, 1995 - 2005)	(1168 projects 2005-2012)
<i>Number</i>	413	324
<i>Fraction</i>	26%	28%
Correctly predicted as unsatisfactory by:		
Full Prediction Model		
<i>Number</i>	191	130
<i>Fraction</i>	46%	40%
Prediction Model with CPIA Only		
<i>Number</i>	148	116
<i>Fraction</i>	36%	36%
Prediction Model with TTL Only		
<i>Number</i>	174	123
<i>Fraction</i>	42%	38%
Prediction Model with CPIA and TTL		
<i>Number</i>	186	135
<i>Fraction</i>	45%	42%

How impressed should we be with this predictive performance? Ideally we would like to perfectly predict both successful and unsuccessful outcomes, but this is not possible. A first benchmark against which to assess the predictive performance of the model is the opposite extreme, in which we just randomly predict that 26 percent of projects are unsuccessful. Under this benchmark, we would correctly predict successful projects 74 percent of the time, and failed projects 26 percent of the time.⁴ The prediction model clearly does better than this benchmark, correctly predicting unsuccessful projects 46 percent of the time, and out-of-sample correctly 40 percent of the time.

A more relevant benchmark is how well this empirical model does relative to early-warning indicators of project outcomes that are available over the course of project implementation. Specifically, can this model do a better job of predicting outcomes than TTLs' own assessments of the project's interim performance, as measured by the ISR-DO ratings? To answer this question, we construct dummy variables capturing the ISR-DO ratings during the first, second, third and fourth

⁴ If predictions are made at random, the probability that a project is successful and is predicted to be successful is p^2 , where p is the success rate of projects. Since only a fraction p of projects actually are successful, the rate of correct predictions of successful projects is $\frac{p^2}{p} = p$. The same argument applies for unsuccessful projects.

quarters of the life of the project. We define project life by two different measures: by cumulative disbursements as a fraction of initial commitments, and by calendar time as a fraction of the planned length.⁵ Specifically, we retrieved the end-of-FY ISR-DO ratings for each year of the life of the project. We then calculated dummy variables equal to one if progress to the DO was rated as satisfactory for each year within the relevant quarters of the life of the project, and zero otherwise.

Figure 2 summarizes how well these ISR-DO ratings predict ultimate project outcomes. The horizontal black line shows the fraction of unsatisfactory projects correctly identified by the prediction model, while the vertical bars show the fraction of unsatisfactory projects flagged as such by the ISR-DO ratings at different points in the life of the project. The early ISR-DO rating in the first quarter of the life of the project is a very poor predictor of unsatisfactory outcomes, capturing only 17 (3) percent in sample with project life defined by disbursements (calendar) time (Figure 2a). The results are similar out of sample, where the ISR-DO ratings correctly predict 21 (4) percent of the unsatisfactory projects in the first quarter of disbursement (calendar) time (Figure 2b).

ISR-DO ratings are better predictors of IEG outcomes later in the life of the project, as measured by both calendar and disbursement time. However, there is still a striking rate of disconnect—roughly half of projects that IEG ultimately will rate as unsatisfactory were rated as satisfactory in the last quarter of their life. More specifically, the ISR-DO rating in the last quarter of project life, as measured by disbursements (calendar time) correctly predict 46 (53) percent of unsatisfactory projects in sample, and 37 (42) percent out of sample. In contrast, the simple prediction model estimated above (represented by the horizontal line in Figures 1), which uses *only* information available at the beginning of the project, does a substantially better job of predicting outcomes throughout the life of projects. It correctly predicts 56 percent of the unsatisfactory projects in sample, and 44 percent out of sample.⁶ Finally, we note that the predictive power of the ISR-DO ratings is better in the first half of project life as measured by disbursements than when the first half is measured in terms of calendar time. One reason for this may be that, since disbursements tend to be backloaded in projects, the first half of project life as measured by disbursements ends nearer to the eventual closing date of the project than the first half as measured by calendar time.

⁵ We observe project ISR-DO ratings only at the end of each fiscal year. So that we can sensibly divide project length into quarters, for this comparison exercise we consider only projects that lasted at least four years.

⁶ Note that the success rates reported in Figure 2 differ from those reported in Table 3. This is because, for the purpose of breaking down results by different quarters, we are restricting the dataset to include only projects at least four years long. Coincidentally the prediction model performs somewhat better in this particular sample.

Figure 2a: Comparison of the in-sample predictive performance during different stages (quarters) of project life.

Figure 2b: Comparison of the out-of-sample predictive performance during different stages (quarters) of project life.

Thus far, the results indicate that the simple prediction model does a better job of predicting ultimate project outcomes than "early" ISR-DO ratings during the first three quarters of the life of the project, and does roughly as well as the last-quarter ISR-DO ratings. This suggests that there may be scope to do better by combining the information in these two alternative leading indicators of project outcomes, particularly to the extent that these two leading indicators identify different sets of projects as potentially problematic. Figure 3 summarizes the contribution of the prediction model and the ISR-DO ratings to the combined predictive power of the two. Specifically, we identify those projects that are correctly identified as unsatisfactory by (i) only the prediction model, (ii) only the ISR-DO rating, and (iii) both the prediction model and the ISR-DO rating. We do this for each quarter of the life of the project, again distinguishing calendar and disbursement time.

Combining these information sources provides a more robust way of predicting unsatisfactory project outcomes as rated by IEG – together they successfully predict between 45 and 66 percent of unsatisfactory projects. Figure 3 also highlights the value-added of the prediction model over the ISR-DO ratings. For instance, during the first quarter of disbursement life (Figure 3b), in addition to the 7 percent of unsatisfactory projects correctly identified by ISR-DO ratings alone (red bars), the prediction model uniquely identifies another 30 percent. Finally, another 14 percent are identified by both the ISR-DO ratings and the prediction model (green bars). The added value of the prediction model is consistently larger than that of the ISR-DO ratings, with the most significant value added in the first quarter of project life as measured by calendar time (see first bar, Figure 3b). In this specific case, the model, correctly and uniquely, identifies 41 percent of the unsatisfactory projects, which is in stark contrast to the 1 percent uniquely identified by the ISR-DO rating. The percentage identified by both measures is just 3 percent of the unsatisfactory projects.

Figure 3a: The added value of the model versus the ISR-DO ratings in terms of predicting unsatisfactory projects, by quarter of project life measured by disbursements

Figure 3b: The added value of the model versus the ISR-DO ratings in terms of predicting unsatisfactory projects, by quarter of project life measured by time.

In this discussion we have emphasized the predictive power of the ISR-DO rating because it conceptually corresponds most closely the assessment of the development objective of the project that IEG performs in its evaluations. However, TTLs are also asked to flag difficulties in project implementation using the ISR-IP flag in each ISR. It is possible that this flag brings additional information about eventual project outcome ratings by IEG. To investigate this possibility, we reproduced the analysis in Figure 2 and 3, for two different variants. In the first, we simply replace the ISR-DO rating with the ISR-IP rating. In this case, we find that the predictive power of the ISR-IP rating is slightly better than that of the ISR-DO rating. For example, in the out-of-sample predictions in Figure 2b, the ISR-IP flag in the last quarter of project life as measured by disbursements (time) correctly predicts 48 percent (44 percent) of unsuccessful project, while for the ISR-DO flag the corresponding figure is 37 percent (42 percent). In the second variant, we constructed a hybrid of the two flags which took the value 1 if either the ISR-DO flag or the ISR-IP flag were raised. In the out-of-sample predictions in Figure 2b, the hybrid flag in the last quarter of project life as measured by disbursements (time) correctly predicts 42 percent (53 percent) of unsuccessful projects. However, our conclusion that the prediction model does a better job of predicting project outcomes relative to the ISR flags in the first half of project life remains true for either of these two variants.

6. Integration with Regular Portfolio Monitoring

In this section we illustrate how these predictions can be integrated into regular monitoring of the World Bank's portfolio of active projects, in order to better identify projects at risk of not having satisfactory outcomes. We first re-estimate the basic prediction model using data from all evaluated projects 1995-2012. The results are shown in Table 4. The results are broadly similar to those in Table 2, except that project length falls just short of significance at the 5 percent level, while preparation time is now significant. We then apply the coefficient estimates from this model to generate predicted probabilities of unsatisfactory project outcomes in the active portfolio.⁷ We calibrate these predictions

⁷ The active portfolio is defined by the set of operations--financed by IBRD, IDA, full sized GEF (>\$1 m), large Recipient Executed Trust Funds (> \$5 m), Special Financing, and Montreal Protocol TF--that were open on June 30, 2013. Some projects in the active portfolio have missing data for some of the variables included in the prediction model. In order to be able to generate predicted outcomes for all projects, we impute these using the sample means of the same variable for all other projects. This affects relative few projects, except in the case of (i) the TTL record variable, which is not available for about 10 percent of projects (reflecting the entry of "new" TTLs into the portfolio, who do not have any IEG evaluations of their past projects in the historical data on which the TTL record variable is based), and (ii) time from approval to effectiveness, which is not available for the roughly 20 percent of projects in the active portfolio that have been approved but are not yet effective

to generate a 27 percent rate of unsatisfactory projects, in line with historical experience over the 1995-2012 period, in which the model was estimated.

Table 4: Basic Prediction Model Estimated Using All Projects 1995-2012

Probit regression	Number of obs	=	2729
	LR chi2(6)	=	197.77
	Prob > chi2	=	0.0000
Log likelihood = -1492.9973	Pseudo R2	=	0.0621

iegsat	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lntotalcom	.0436975	.0234646	1.86	0.063	-.0022923	.0896874
preptime	-.0036075	.0016079	-2.24	0.025	-.006759	-.0004561
effdelay	-.0040676	.0056336	-0.72	0.470	-.0151093	.0069741
planlength	-.0031519	.0016926	-1.86	0.063	-.0064693	.0001655
ttlqual	.9578754	.0963074	9.95	0.000	.7691164	1.146634
cpiaav	.3770521	.0575975	6.55	0.000	.2641632	.4899411
_cons	-1.175633	.226064	-5.20	0.000	-1.61871	-.7325552

The distribution of predicted unsatisfactory projects across regions and different stages of project life is summarized in Tables 5 and 6. The first column of Table 5 contrasts the prediction model with the ISR-DO ratings, pooling all 1591 projects. By assumption, the prediction model generates a 27 percent rate of unsatisfactory projects, in line with the historical data. This, however, is much higher than what is reflected in the ISR-DO ratings, which rate only 13 percent of projects as making unsatisfactory progress towards their development objectives. The predicted rate of unsatisfactory projects varies substantially across regions, ranging from 15 percent in ECA to 41 percent in AFR and 43 percent in MNA. In all regions other than ECA, the prediction model suggests a higher rate of unsatisfactory projects than the ISR-DO ratings, and sometimes by a substantial margin. The largest disconnect between the two is for projects in AFR -- while the ISR-DO ratings predict that only 10 percent of the Bank's 483 active projects in the AFR region will lead to unsatisfactory outcomes, the corresponding prediction provided by the model is 41 percent.

It is also interesting to compare the predicted rate of unsatisfactory projects with the historical averages by region. If the prediction model (which is estimated pooling data for projects in all regions) is stable across regions, and if the distribution of explanatory variables in the prediction model is reasonably stable over time, we should expect to see regionally-disaggregated predictions that are similar to historical average rates of project success. This is reasonably true for most regions. For example, in AFR the historical rate of unsatisfactory projects is 38 percent while the prediction model

generates 41 percent. At the opposite extreme, the historical rate of unsatisfactory projects in EAP is 18 percent and the model generates 17 percent. Only one region, MNA, stands out as having a large gap between the two: the historical rate of unsatisfactory projects is 31 percent while the model predicts 43 percent.⁸ However, a crucial feature of the prediction model is that it allows us to identify specific projects in specific regions at greater risk of ultimately being unsatisfactory -- something that cannot be done by just looking at past rates of project success by region.

Table 5: Predictions for projects in the active portfolio by Region.

	ALL	AFR	EAP	ECA	LCR	MNA	SAR
Historical Project Performance							
% Rated Unsatisfactory by IEG, 1995-2012:	27%	38%	18%	20%	22%	31%	27%
Active Portfolio							
# Projects:	1591	483	291	206	265	120	222
% Rated Unsatisfactory by ISR-DO:	13%	10%	14%	17%	14%	21%	10%
% Rated Unsatisfactory by Prediction Model:	27%	41%	17%	15%	20%	43%	22%
of which: also Rated Unsatisfactory by ISR-DO	5%	6%	3%	3%	5%	11%	3%

Table 6 summarizes the predictions by project life. By construction, the model predicts a similar proportion of unsatisfactory projects around 27 percent across all stages of project life. This is because the inputs into the prediction model do not vary systematically over the life of the project. The key point however is that these predictions are substantially different from those provided by the ISR-DO ratings. For example, only 16 percent of projects having disbursed 25 percent or less are rated by TTLs as having unsatisfactory ISR-DO ratings, even though historical experience suggests that around 27 percent of them will ultimately be unsatisfactory. This disconnect is even starker for projects in the first quarter of their planned lifespan, where only 3 percent are flagged as unsatisfactory with the ISR-DO ratings.

⁸ It is of course possible to use region-specific thresholds to calibrate the fraction of projects identified as at risk of being unsatisfactory by the prediction model, in order to eliminate this potentially undesirable feature of the results.

Table 6: Predictions for projects in the active portfolio by project life.

	Project Life by							
	Disbursements				Calendar Time			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
# Projects:	689	250	243	405	261	288	308	689
% Rated Unsatisfactory by ISR-DO:	16%	16%	13%	6%	3%	10%	19%	16%
% Rated Unsatisfactory by Prediction Model:	27%	28%	27%	26%	22%	31%	23%	28%
of which: also Rated Unsatisfactory by ISR-DO	6%	8%	4%	2%	2%	6%	5%	6%

7. Summary and Conclusions

This paper has shown that the empirical relationship between project characteristics and project outcomes can serve as a useful tool for identifying likely project outcome ratings. This prediction model is of course far from perfect: in the historical data it fails to predict all unsatisfactory projects correctly, and moreover incorrectly identifies some ultimately satisfactory projects as being at risk of unsatisfactory outcomes. However, it does provide two crucial features. First, it generates predicted outcomes for each individual project in the active portfolio that arguably add value to those embodied in the ISR-DO ratings. Second, by flagging individual projects at risk of unsatisfactory outcomes based on their observed characteristics, including country CPIA ratings and the track record of the TTLs, the prediction model provides a useful tool for World Bank management to identify projects at early stages of implementation where proactive interventions may be effective in leading to better project outcomes.

References

- [1] Deininger, Klaus, Lyn Squire, and Swati Basu (1998). "Does Economic Analysis Improve the Quality of Foreign Assistance?". *World Bank Economic Review*. 12(3):385-418.
- [2] Dollar, David and Jakob Svensson (2000). "What Explains the Success and Failure of Structural Adjustment Programs?". *The Economic Journal*. 110():894-917.
- [3] Kilby, Christopher (2000). "Supervision and Performance: The Case of World Bank Projects". *Journal of Development Economics*. 62: 233-259.
- [4] Chauvet, Lisa, Paul Collier, and Andreas Fuster (2006). "Supervision and Project Performance: A Principal-Agent Approach". Manuscript, DIAL.

[5] Chauvet, Lisa, Paul Collier, and Margeurite Duponchel (2010). "What Explains Aid Project Success in Post-Conflict Situations?". World Bank Policy Research Working Paper No. 5418.

[6] 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.

[7] Guillaumont, Patrick and Rachid Laajaj (2006). "When Instability Increases the Effectiveness of Aid Projects". World Bank Policy Research Working Paper No. 4034.

[8] Dreher, Axel, Stephan Klasen, James Raymond Vreeland, and Eric Werker (2010). "The Costs of Favouritism: Is Politically-Driven Aid Less Effective?". CESIFO Working Paper No. 2993.

[9] World Bank (2012) "Delivering Results by Enhancing Our Focus on Quality", Operations Policy and Country Services

[10] Independent Evaluation Group (IEG), World Bank, 2009. *Annual Review of Development Effectiveness*. Washington, DC.