Policy Research Working Paper 7800

Predicting Project Outcomes

A Simple Methodology for Predictions Based on Project Ratings

Marc Blanc
Talib Esmail
Caroline Mascarell
Rukshan Rodriguez

WORLD BANK GROUP
East Asia and the Pacific Region
Development Effectiveness Unit
August 2016

Policy Research Working Paper 7800

Abstract

The downgrading of projects at the closing from moderately satisfactory to moderately unsatisfactory has been a persistent problem in the World Bank and a particular problem in the World Bank's East Asia and Pacific region since 2012. Through analysis of the projects that exited the East Asia and Pacific region's portfolio in fiscal years 2012 and 2013, this paper derives a prediction model based on ratings for implementation progress and achievement of development objectives during project supervision. The model, used in combination with other indicators of project progress toward outcomes, appears to improve on existing methods for assessing the downgrade risk.

This paper is a product of the Development Effectiveness Unit, East Asia and the Pacific Region. 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 tesmail@worldbank.org, cmascarell@worldbank.org, and mblanc@worldbank.org.

Produced by the Research Support Team

Predicting Project Outcomes

A Simple Methodology for Predictions Based on Project Ratings

Marc Blanc

Talib Esmail

Caroline Mascarell

Rukshan Rodriguez

Acknowledgments

This paper was prepared with the assistance of Brigitte Duces, Carlos Elbirt, Vilija Kostelnickiene, Edgar Molina, and Katia Nemes. William B. Hurlbut (consultant) edited the paper.

I. Introduction

In its *Results and Performance of the World Bank Group 2015*, the Independent Evaluation Group (IEG) reported a decline in the outcome ratings of World Bank investment projects that was particularly strong in the portfolio of the East Asia and Pacific (EAP) region. Understandably, this triggered concern within the EAP vice-presidential unit (VPU) about the efficacy of existing systems to provide warnings about the potential that a project will fail to deliver its intended results.

In an attempt to discover what went wrong, the VPU undertook an analysis of the projects that contributed to the decline in ratings. That analysis led to the discovery of a new method for predicting when investment projects that, under current monitoring regimes, appear to be headed toward a moderately satisfactory rating for outcome, but may be at risk of a downgrade on completion. This paper provides some background on the decline in ratings and current monitoring systems, details the portfolio analysis, describes the prediction model that evolved from that analysis, and reports on validation of the model through subsequent testing.

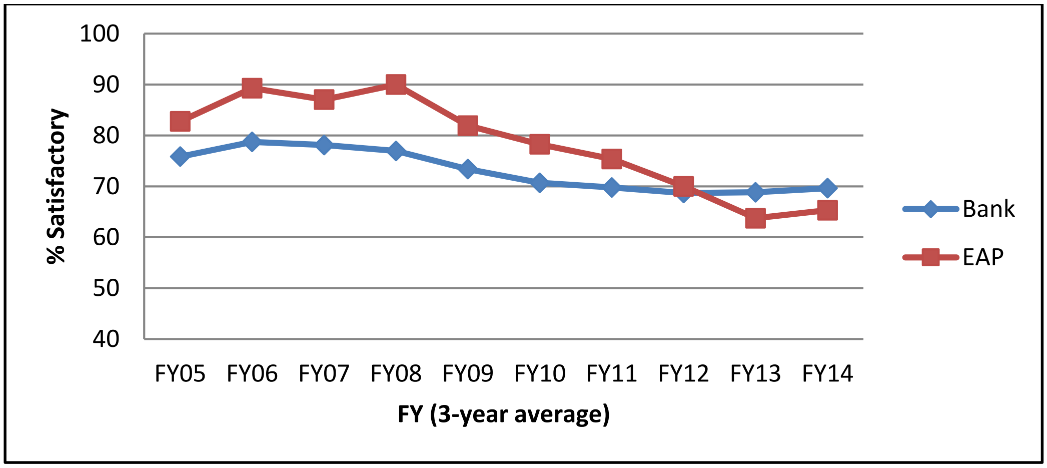
Reasons for Concern about Outcome Ratings in EAP

Outcome Ratings Have Declined

Outcome ratings for IBRD and IDA investment projects, based on the three-year average reported by IEG in its annual *Results and Performance* reports, have been declining across the Bank since 2006.[1] The ratings in EAP have followed the broader trend, but on a somewhat steeper trajectory and in the 2012 fiscal year the ratings for the region dropped below the Bank-wide average. A significant uptick in the FY14 outcome ratings for both the Bank and the region[2] explains the upward shift in the three-year average for the last year in the series (Figure 1).

Figure 1. Outcome Ratings, Bank versus EAP

Three-year average for investment project financing for IDA and IBRD



[1] These data reflect IEG results as of December 31, 2015. The three-year average for the percent of satisfactory outcome ratings (i.e. ratings of moderately satisfactory or better) is based on 115 FY12-14 EAP projects rated covering 87 percent of the FY12-14 exits. Out of 42 FY14 exits 35 projects had been rated.

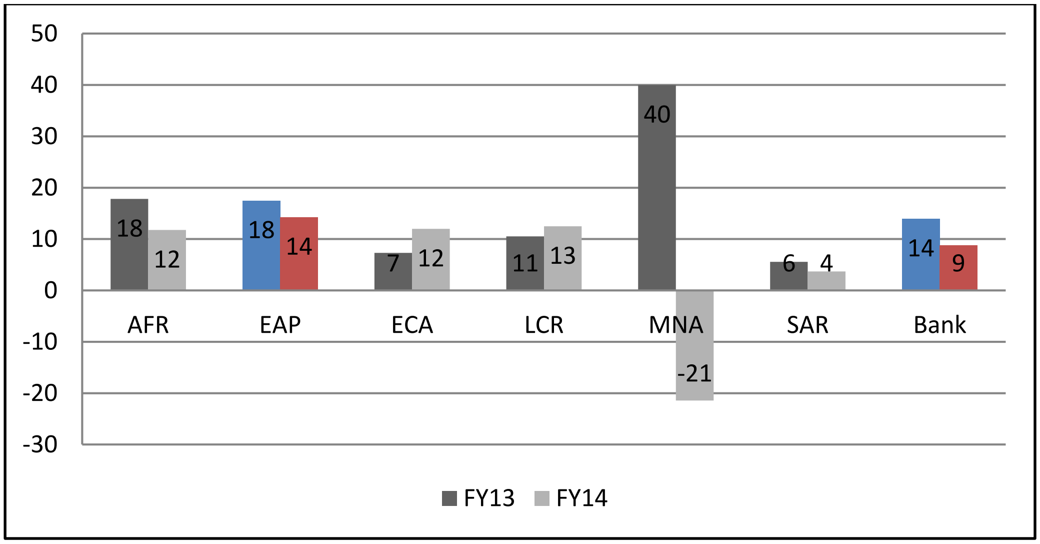
[2] When ratings are weighted by net commitment amounts, the percentages of satisfactory ratings are higher, indicating that larger projects appear to have been more successful than smaller ones over the period considered. However, the main concern is with the declining number of successful projects.

Regression analyses of project performance at exit (outcome ratings) consistently have shown a high degree of correlation with quality at entry and quality of supervision. While this is to be expected since these ratings are all given at exit by the same evaluator, it is still useful to observe that quality at entry and the quality of supervision ratings, for both the Bank and EAP, have tracked the declining outcome ratings.

Net Disconnect and the Candor Gap Have Increased

In addition to the decline in satisfactory outcome ratings, the two measures of realism in self-evaluation ratings have increased. The net disconnect is the difference between the percentage of projects IEG rates unsatisfactory for development outcome and the percentage the region's final Implementation Status and Results (ISR) reports rates unsatisfactory for achieving their development objectives. Although net disconnects have declined from their 18 percent level in FY13 they have remained larger than the Bank average, and have been associated with the region's declining outcome in years prior to FY14 (Figure 2). While there are substantial variations from year to year and among country management units, there is a strong relationship between the low and decreasing ratings for monitoring and evaluation quality and the substantial disconnect. Projects with unclear and weak results frameworks are more difficult to evaluate, giving rise to the disconnect.

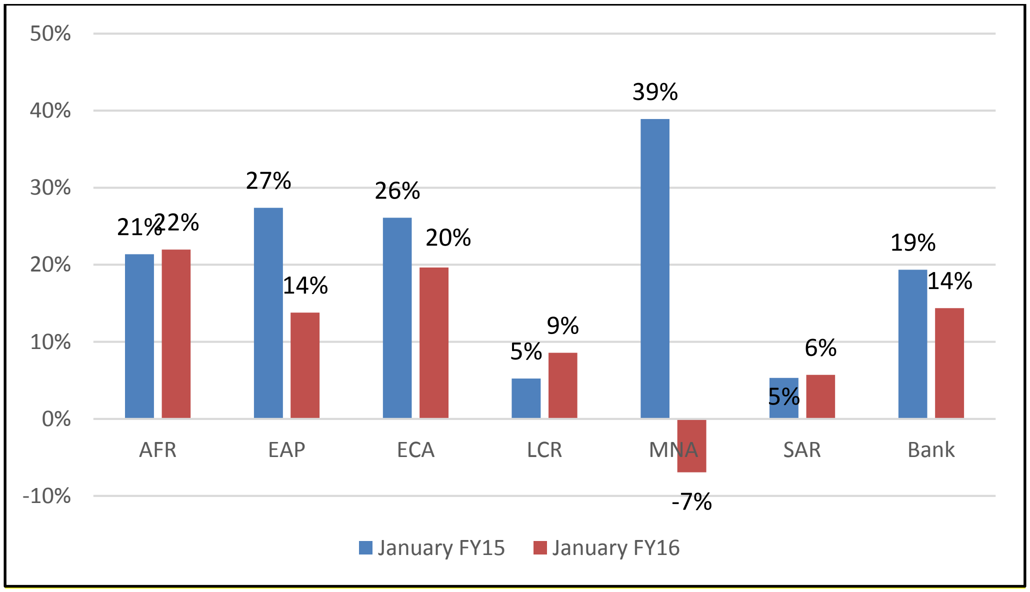
Figure 2. FY13 versus FY14 Net Disconnect by Region (IBRD and IDA)



Note: 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 Northern Africa, SAR = South Asia

The candor gap is the difference between the percentage of projects with satisfactory development outcome ratings in the active portfolio for a fiscal year (including both investment project financing and development policy financing) and IEG ratings of satisfactory outcome based on projects evaluated over the past 18 months as of the date of the data download. The candor gap for EAP was higher than the average for the Bank and all other regions (except the Middle East and Northern Africa region) up to January of fiscal 2015. However, since the gap closely tracks recent IEG project outcome ratings, it has improved in line with the improvements in FY14 outcome ratings for the region, and reached the Bank average of 14 percent as of January 2016 (Figure 3).[3]

Figure 3. Candor Gap by Region (IBRD and IDA)



Note: 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 Northern Africa, SAR = South Asia

Shortcomings in Existing Performance Monitoring

While the candor gap is useful for identifying the scale of the performance problem and assessing the level of over-optimism among task teams, it does not help to identify which projects may need attention during implementation to reduce the likelihood of a less than satisfactory rating upon closing. For that purpose, the Bank uses two other tools: disbursement tracking and ratings for implementation progress (IP) and development outcome (DO) in ISR reports that are prepared about every six months during the life of the project. Both methods have limitations.

Disbursement tracking underestimates projects likely to be rated moderately unsatisfactory

The Bank's disbursement monitoring systems track "disbursement ratios" for countries and regions as well as tracking slow-disbursing projects. Currently, a monthly senior management meeting identifies and reviews all IBRD and IDA projects more than 5 years old and with undisbursed balances greater than 60 percent. As of the end of 2015, EAP had only three slow-disbursing projects and only 5 percent of EAP projects were delayed by 24 months or more. This substantially underestimates the percentage projects likely to be rated moderately unsatisfactory or lower at exit, which is currently on the order of 30 percent. Moreover, regression studies carried out by others (see bibliography) have failed to demonstrate a strong negative correlation between the disbursement delays and project outcome.

[3] If there was no lack of candor or over optimism in reporting by task team leaders (TTLs), one could argue that the underestimation and overestimation of the project risks would cancel each other out, leading to a non-biased estimate. However, this is not the case as easily demonstrated by the very low amount of negative disconnects when looking at the difference between the IEG ratings and the final ISR rating, i.e., when IEG rates a project moderately satisfactory or higher that had been rated moderately unsatisfactory or lower in the last ISR.

ISR ratings for IP and DO also underestimate projects likely to be rated moderately unsatisfactory

During routine project supervision projects are rated on their IP and DO. A moderately unsatisfactory rating is given to projects that fall short on performance standards and are therefore likely to fail at meeting their development outcome at exit. These projects are then referred to as "Problem Projects." Table 1 shows the percentages of projects rated moderately unsatisfactory for IP, DO, or both as of the end of 2015. These percentages are believed to be substantially below the failure rates of these projects at exit. As previous studies have shown, this is a very consistent and stable occurrence across both time and institutional levels, since it is rooted in the over-optimism or lack of candor of task team leaders when they self-report on their project in the ISR report.

Table 1. Percentage of Moderately Unsatisfactory or Lower Ratings for DO and IP, all Investment Projects as of December 31, 2015

	DO	IP	DO or IP
Bank	15%	20%	21%
EAP	15%	21%	22%

II. Existing Approaches to Predicting Performance

As the preceding discussion shows, the current tracking mechanisms are not very effective at predicting whether a project rated moderately satisfactory for most of its life will be downgraded on completion. But those are not the only ways that the Bank tries to manage its risks.

Risk at entry – difficult to ensure update during implementation

One method for managing risk identifies and defines the level of risk at entry, that is, during the preparation and approval phase of the project, using relatively well defined criteria and methodologies. This approach has been useful at the preparation stage to identify comprehensively the risk profile of the project, and thus the type of processing and level of resources required. It has proved much less useful after Board approval, probably because it is difficult to get the task team leader to regularly and precisely update this risk rating during project supervision.

Risk under implementation – a system that has resulted in a perverse incentive to manage flags

Another approach has identified a risk profile of the project under implementation based on flags, or warning indicators. These have included self-reported indicators, such as for DO and IP, as well as system-based indicators, such as disbursement delay calculated by comparing the actual and the initial or revised disbursement schedules. The full set of indicators, currently numbering about ten, are being input in the ISR report and are tracked in the project monitoring system; several of them are also monitored at the corporate level (for example in the Corporate Scorecard, memoranda of understanding, and senior management monthly monitoring meetings).

Until about 2007 the flag system (then consisting of 12 flags) was used to identify and track "potential problem projects" that were not rated moderately unsatisfactory or lower for IP or DO during implementation. This led some teams to avoid the appearance of failure by finding ways to avoid rating 3 flags as less than satisfactory. This undercut the ability of the system to detect problem projects and the system was dropped, though 10 of the flags remain in use.

Recent studies of risk prediction identified significant predictors of project outcomes – but they still only explain a small percentage of the downgrades

Two recent studies attempted to identify factors contributing to the success and failure of investment projects, and to build prediction models for the outcome ratings of those projects. In the first, Denizer, Kaufmann, and Kraay (2011) analyzed correlations between project outcomes and project characteristics, such as size and duration, as well as country variables and project variables corresponding to the 12 original warning indicators used in the implementation monitoring systems. In the second study, Geli, Kraay, and Nobakht (2014) built on the findings of the previous study and designed an outcome prediction model using a combination of country and project variables, and comparing it with the prediction value of the IP and DO ratings as the project matures.

Denizer, Kaufmann, and Kraay used data on more than 6,000 World Bank-financed projects undertaken in 130 countries since the 1970s to develop their model. Their major findings were:

- Eighty percent of the variation in project outcomes occurs *within* countries rather than *between* countries, and is therefore accounted for by *project* variables rather than *country* variables.
- Country Policy and Institutional Assessment (CPIA) scores account for 40 percent of the between-country variations and therefore for 8 percent of the variation in outcome. It is among the factors most strongly correlated with project outcome.
- Project variables used in the regression analysis, comprising both characteristics (such as size, duration, and preparation costs) and early flags, overall explain only 6 percent of the 80 percent variation, or 5 percent in the project outcome.
- The quality of the task team leader is strongly correlated to the project outcome, with a contribution about equal to that of the CPIA.
- Among the most significant results of individual partial correlation of outcome with project variables were project size, duration, and preparation and supervision costs, which were negatively correlated with outcome; early IP/DO flags, which were negatively and very significantly correlated with the ultimate outcome; and the early monitoring and evaluation (M&E) flag and sometimes the project management flag, which were also significantly and negatively correlated with outcome.
- No evidence was found that disbursement delays are significantly correlated with outcomes.

Geli, Kraay, and Nobakht, for their prediction model, used all investment operations between 1995 and 2005,[4] and included correlates for size, preparation time, elapsed time between approval and effectiveness, initially planned project length, task team leader track record, and CPIA score. Their major findings were:

- The main correlates with outcome were task team leader track record and CPIA score; preparation time and effectiveness delay were negatively correlated but not significantly so; project size not significantly so.
- Application of the full model with all the variables correctly predicted about 40 percent of the outcomes, but the same result was obtained with only the CPIA and the task team leader track record variables.

[4] After defining the model using the sample of 1,561 projects, with closing dates between 1995 and 2005, the model was applied to a second "out of sample" set of 1,168 projects closing from 2005 to 2012. Results between the two sets were consistent.

- Only in the last quarter of the project's life (measured by time or by disbursement amount) does the DO rating reach the prediction power of the model using only the CPIA and task team leader quality variables. Until then, it is lower than or equal to 20 percent.
- For EAP, the predictive performance of the ISR DO rating was even lower.
- Combining the first model with the ISR DO rating during each quarter of the life of the project leads to a predictive power of slightly above 60 percent in the last quarter.
- Using the IP rating instead of the DO rating improves the prediction model, and using the two together further enhanced the results.
- Applying the model to the EAP portfolio as of July 1, 2014, yielded a predicted rate of unsatisfactory outcome of 17 percent.

The two studies clarify the relative importance of country and project variables, including flags and other indicators for predicting project outcome, but they have several limitations. First, since the model significantly underestimates current IEG rates of unsatisfactory outcomes for EAP, now close to 30 percent, it will need to be re-run and calibrated using more recent data. Second, the quality of the task team leader is a difficult, sensitive, and controversial variable to estimate fairly and regularly. Third, the model still only predicts about 40 percent of the unsatisfactory outcomes.

III. Foundation for a New Outcome Prediction Tool for Moderately Satisfactory (or Higher) Rated Projects

Given the limitations of existing monitoring tools and predictive models, it would be advantageous to find a simpler approach that could yield similar or better results, by using real-time data, such as the IP and DO ratings, even if it meant moving away from a pure prediction model, i.e., a model that would mostly include variables already known at the project start. Hence, EAPDE conducted a thorough analysis of the projects rated less than moderately satisfactory and particularly disconnects in search of a tool that would predict whether a project rated moderately satisfactory would be downgraded on completion.

Methodology for analysis

Since the review started out of concern about the sharp decline in IEG rates of satisfactory outcomes in the latest set of projects reviewed, it was logical to use the set of projects that exited in FY12 and FY13 for analysis. Since the focus was on identifying factors and patterns contributing to the increase in unsatisfactory outcomes, the review concentrated on projects with unsatisfactory ratings, particularly those with disconnects.

The convenience sample included all 62 investment projects that exited in FY12 and FY13 with ICR reviews posted on the IEG site as of November 15, 2014. In accordance with IEG review policy, the sample included all projects financed by IBRD or IDA as well as Global Environment Facility, State and Peace-Building Fund, and recipient-executed trust funding with commitment amounts larger than $5 million.

The IEG ICR reviews were first examined to identify patterns in the country and network or sector breakdown. Outcome results broken down by country were compared with corresponding CPIA figures and with candor gaps and recent series of ratings (Annex 1, Table 1). Since IEG uses four dimensions to assess the outcome rating of a project,[5] the relative impact of each dimension on the outcome rating was analyzed for the projects rated moderately unsatisfactory or lower. The frequency and impact of restructuring were also reviewed.

Next, for each of the projects rated moderately unsatisfactory or lower the incidence of warning indicators or flags collected during their lifetime was noted and recorded. Special attention was given to the IP and DO ratings, but the incidence was also calculated for project management, procurement, M&E, financial management, safeguards, counterpart funding, slow disbursements,[6] legal covenants, effectiveness delay,[7] problem project, and long-term risk.[8] Incidence results were then analyzed and a simple set of rules, derived by inspection, that allowed to most closely match in the sample the percentages of projects that IEG had rated moderately unsatisfactory or lower for their outcomes.

The set of rules was then applied to the EAP FY14 active portfolio. This identified 106 projects that were at risk of not meeting their objectives. A first subset of 42 projects under implementation was then reviewed using the IEG ICR review methodology to identify potential weaknesses.

Results of analysis

The country breakdown of outcome ratings for the 62 projects in the sample is shown in Table 2, along with the corresponding percentage of moderately unsatisfactory or lower in the candor gap set as of the date of the review.

Table 2. Country Distribution of Rated Projects

	China	**Mongolia**	**Indonesia**	**Philippines**	**Vietnam**	**Cambodia**	**Lao PDR**	**Thailand**	**Papua New**	**Timor-Leste**	**Solomon Islands**	**Total region**
#of projects reviewed	12	5	11	6	10	6	4	1	1	4	1	**62**
% MS+	67	60	36	17	70	67	50	0	100	0	100	**52%**
% MU-	33	40	64	83	30	33	50	100	0	100	0	**48%**
% MU- in candor gap set	24	25	37	75	25	50	20	100	N/A	100	N/A	**38%**
Low CPIA countries % MU-						33	50		0	100	0	**50%**
Low CPIA countries: CPIA rating as of 11/2014						3.6	2.8		3.4	2.3	3.3	

[5] The four dimensions are: relevance of objectives, relevance of design, efficacy, and efficiency.

[6] A slow disbursement flag is given to projects when the delay is 24 months or higher. The delay is calculated on based on the initial or "revised" disbursement schedule for the project. The definition of revised has varied: it had been meant to be "officially revised," i.e., sanctioned by a restructuring for many years, but the definition has been relaxed recently to include informal revision as well.

[7] The effectiveness flag is given when the elapsed time between Board approval and effectiveness is more than nine months for investment and three months for emergency operations. It is turned off three years after Board approval.

[8] This flag is no longer officially tracked but is still available in the system. It is given if the project has been rated moderately unsatisfactory or lower for IP or DO for any 24 months cumulative during the life of the project. The flag is removed when the project has been rated moderately satisfactory or higher for IP and DO for the previous 24 months. This flag is currently used to identify the "non-proactive" projects, which are then subjected to special reviews.

The following observations can be made:

- The variations by country of the MU- and MS+ percentages are very large because of the small sample reviewed, which for several countries include only one project.
- The sample shows a higher percentage of MU- than the candor gap reference set, but the correspondence with the percentages of the candor set is significant. This confirms the relative stability of the "country record" outcomes over a period of a few years.
- The percentage of MU- for the five countries that had a CPIA lower than 3 at the time the project was implemented is 50 percent, i.e., about equal to the percentage of the full sample. This is moot for confirming the relationship between the CPIA rating and the outcome rating, but can be expected given the small sample size, and the fact that a couple of countries had already improved their CPIA at the time the projects exited (Cambodia and Papua New Guinea).

The network/sector breakdown of the outcome and disconnect ratings for the sample reviewed as compared with a three-year FY11-13 moderately satisfactory or higher average is shown in Table 3.

Table 3. Network/Sector Distribution of Rated Projects

	SDN	HDN	PREM	FPD	Total
# of projects reviewed within 62 FY12-13	42	16	2	2	62
% MS+ FY11-13	68	73	69	100	
% MS+	48	69	0	50	52
%MU-	52	31	100	50	48
Of which disconnect %	31	19	0	0	26

Notes: SDN = Social Development Network; HDN = Human Development Network; PREM = Poverty Reduction and Economic Management; FPD = Financial and Private Sector Development

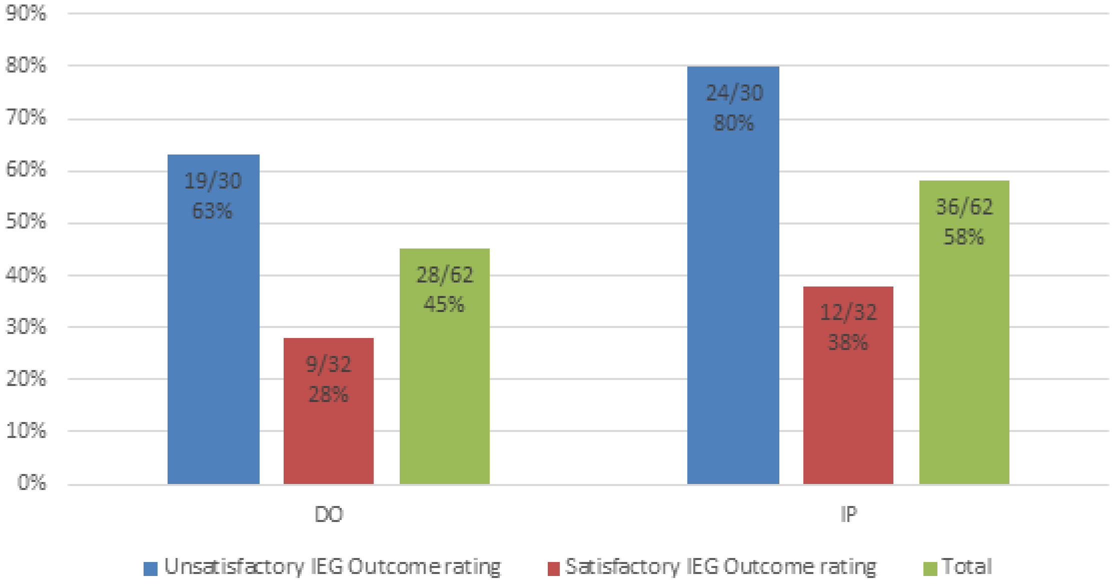
- The percentages of MS+ for SDN and HDN are roughly consistent, but they are not for PREM and FPD because of the small sample.

The results of the incidence analysis for DO and IP ratings are summarized in Table 4 and Figures 4 and 5.

Table 4. Incidence of Outcome Ratings and Disconnects

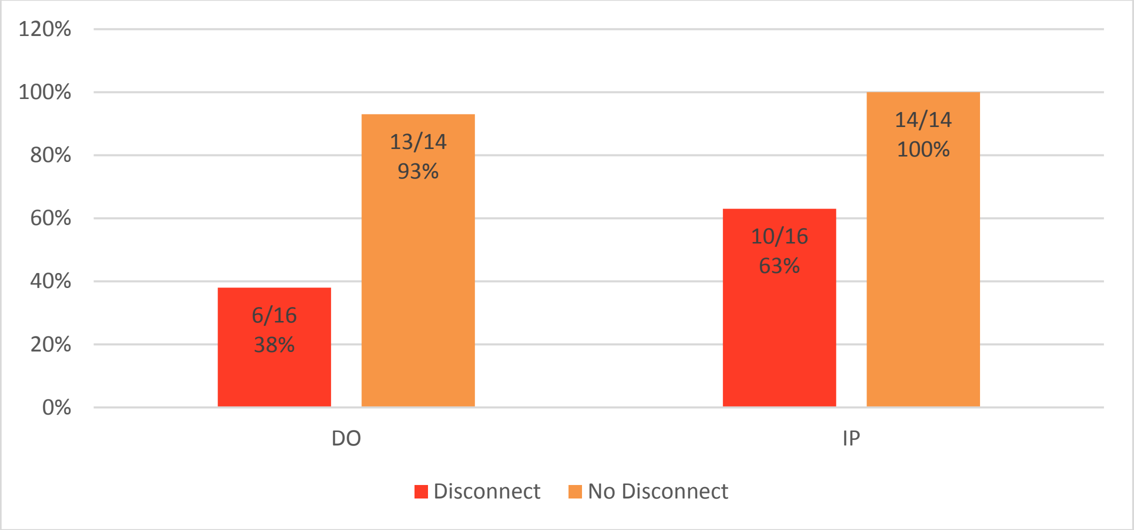
	MU- DO ratings		MU- IP ratings	
	Last DO rated MU-	DO Ever rated MU-?	Last IP rated MU-	IP Ever rated MU-?
16 Disconnects 26%	0 (0%)	6 (38%)	0 (0%)	10 (63%)
14 MU- outcome & No disconnect 23%	13 (93%)	13 (93%)	12 (86%)	14 (100%)
Sub-Total 30 MU- rated 48%	13 (43%)	19 (63%)	12 (40%)	24 (80%)
Sub-total 32 MS+ rated 52%	0 (0%)	9 (28%)	0 (0%)	12 (38%)
Total 62 sample 100%	13 (21%)	28 (45%)	12 19%	36 (58%)

Figure 4. Percentage of Projects That Ever Had an Unsatisfactory Rating for IP or DO During the Project Life



- The last ISR DO/IP rating at closure matches the IEG rating for all projects in the sample except for disconnects. This is clearly consistent with the statistical observation that the predictive performance of the IP/DO rating improves as the project matures.
- Twenty-eight percent of the MS+ IEG-rated projects in the sample have ever been rated MU- for DO, while 63 percent of the IEG-rated MU- have been rated MU- at some stage.
- For IP MU- percentage ratings the corresponding incidence figures are 38 percent for the MS+ IEG-rated, compared with 80 percent for the MU- IEG-rated, factoring in 63 percent for disconnects.
- Among the 16 disconnects, 6 were ever rated MU- for DO (38 percent), compared with only 9 of the 32 MS+ outcome (28 percent).
- Out of the 28 projects ever rated MU- for DO, only 9 ended up having an MS+ outcome. In other words, those projects that have ever had one MU- rating for DO have about two of three chances of exiting with an MU- rating.
- Similarly, out of the 36 projects ever rated MU- for IP, only 12 ended up having an MS+ outcome.
- A significant percentage of unsatisfactory projects with disconnect were rated MU- for their IP or DO at some stage in their project life.
- These percentages are significantly higher than those observed for MS+ projects.

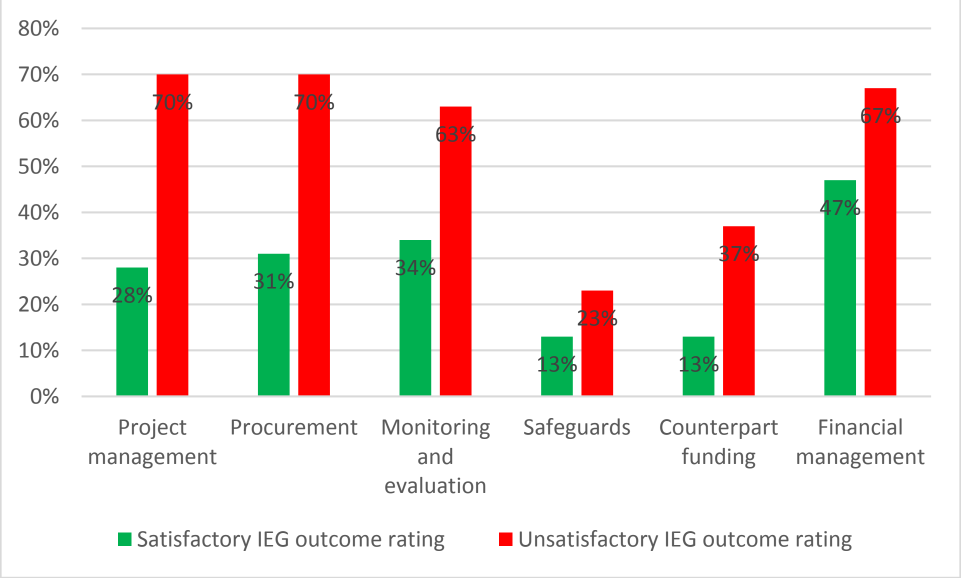
Figure 5. Percentage of Unsatisfactory Projects That Ever Had an Unsatisfactory Rating: Disconnect versus No Disconnect



Note: One of the 14 projects with unsatisfactory IEG rating had never received an MU- DO rating, but was rated MU- at exit for IP. It was considered a non-disconnect in this analysis.

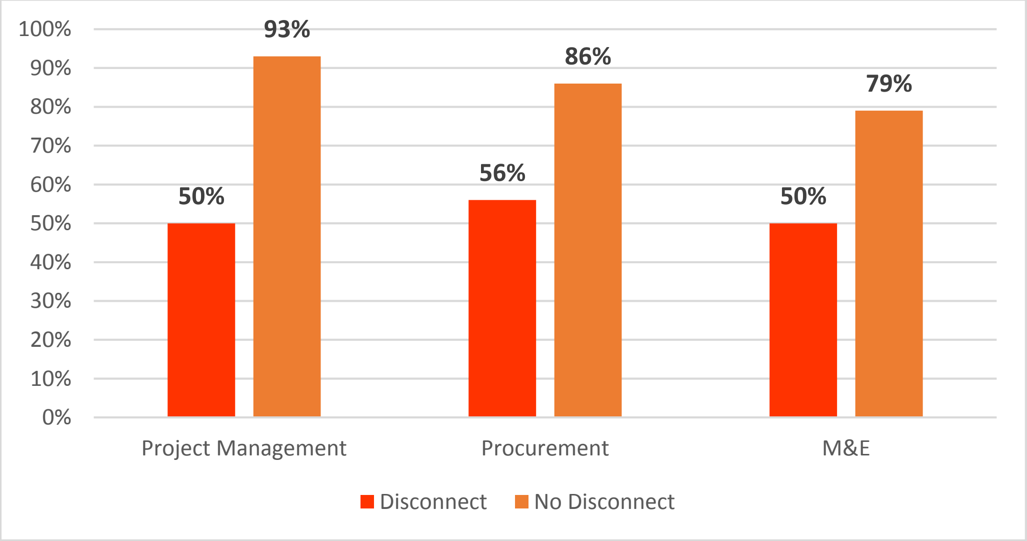
The results of the incidence analysis for six indicators—project management, procurement, M&E, safeguards, counterpart funding, and financial management—are summarized in Figures 6 and 7.

Figure 6. Percentage of Projects That Have Ever Had an Unsatisfactory Rating During the Life of the Project



The incidence of three of these indicators is significantly different for projects with disconnect than it is for those without, as Figure 7 shows.

Figure 7. Percentage of Unsatisfactory Projects That Have Ever Had an Unsatisfactory Rating During the Project Life



- A high percentage of MU-rated projects with no disconnect had an unsatisfactory rating for project management, procurement, and M&E, at some point in the project life.
- A significant number of MU- projects with disconnect received an unsatisfactory rating for project management, procurement, and M&E, at some point in the project life.
- However, these last percentages are significantly higher than those observed for MS+ projects as shown in Figure 6.

The above results are fully consistent with previous regression analyses (Denizer, Kaufmann, and Kraay 2011 and Geli, Kraay, and Nobakht 2014), which have shown a strong correlation between outcomes and the M&E and project management indicators, and to a lesser extent the procurement indicator. Poor M&E, in particular, has been the most consistently and strongly correlated to an unsatisfactory outcome.

Even though safeguards, counterpart funding, and financial management do not show as strong an incidence and correlation as the other indicators, they still capture a high percentage of unsatisfactory projects and show a significantly different incidence between MS+ and MU- rated projects.

In summary, in addition to the IP and DO ratings, which have been shown to be significantly correlated with the final IEG outcome rating of any project, the six analyzed indicators, which represent critical dimensions of an investment project's implementation are critical markers and predictors of a project's success. Under that premise, a trial and error, heuristic search was conducted to identify the combination of these indicators that would provide the best predictor of future successful outcome rating.

The Prediction Model and Verification Testing

The major finding of the analysis is that projects with a relatively high risk of being rated unsatisfactory at exit include two groups. The first is projects that are, at the time of the review, rated moderately unsatisfactory or lower for IP or DO in their current ISR. The other group includes those projects that are not currently rated moderately unsatisfactory or lower for IP or DO but that have had at least three of six indicators rated moderately unsatisfactory or lower at any time during their life, and not necessarily concurrently. The two groups constitute separate modules in the proposed prediction model. The first can be called the problem project module, the second can be called the flag-based module.

Validation of the flag-based module was undertaken first at the level of the sample of 62 reviewed EAP projects, and then at the level of a corresponding Bank-wide sample of 531 projects including all FY12-13 investment project exits rated by IEG as of February 1, 2016.

Applying the flag-based module to the 62 projects in the study sample, and to the 531 in the Bank-wide sample, had an overall ex post prediction rate of 76 percent for the set of EAP projects and 68 percent for the set of Bank-wide projects (Table 5).

Table 5. Ex Post Prediction Rates Using the Flag-Based Module

At least 3 of the 6 flags at any time	EAP	Bank	EAP (#)	Bank (#)
Predicting an unsatisfactory rating	70%	59%	21/30	103/174
Predicting a satisfactory rating	81%	72%	26/32	258/357
Overall prediction rate	76%	68%	47/62	361/531

*EAP sample size is 62
**Bank sample size is 531

The likely reason the ex post prediction rates are lower for the Bank-wide sample is the particularly low performance of the EAP sample in the reviewed period, as well as the low candor level of ratings. With an almost 50 percent unsatisfactory outcome and 26 percent disconnect, the flag-based module is able to identify a significant number of those disconnects.

These ex post prediction rates are obviously higher than those that could be estimated on an actual "running" portfolio for at least two reasons. First, at project closure the last DO rating normally has its highest prediction rate, which is highest when the "disconnect" is the lowest: i.e., lower rate for the EAP sample due to high disconnect of 27 percent and higher rate for the Bank with an overall disconnect of 20 percent. Second, at closure, any project would have received its highest allocation of flags; therefore, the flag-based module would yield a lower prediction when applied to a running portfolio. How much lower cannot be known, as it would depend on the average age of the portfolio and the rating candor in the region. When applied to the actual EAP portfolio the corresponding FY15 and FY16 percentages identified by the flag-based module were 22 percent (66/297) and percent 18 percent (51/284).

Creating a "watch list" for FY15 projects potentially at risk of a moderately unsatisfactory (or lower) outcome rating

The model, consisting of the problem project module and the flag-based module, was applied to the EAP FY15 investment project portfolio of 297 projects active as of July 1, 2014, to identify a watch list of projects at risk of not achieving their outcomes.

The portfolio consists of projects in three age groups: 63 projects had reached their final year of implementation, and were due to close in FY15; 59 new projects approved during FY14 were less than one year old; and 175 projects were older than one year and not due to close during FY15. The application of the flag-based module and problem project module identified 30 projects in the first group, 2 projects in the second group, and 65 projects in the third group.

While the combination of projects with IP or DO rated moderately unsatisfactory or lower and those with three out of six flags may be a reasonable predictor for well-established projects, application of these two rules alone would not select recently approved projects, which are seldom rated moderately unsatisfactory or lower for any indicator. Hence, to complete the prediction model, and to potentially increase the prediction rate, it seems appropriate to include a third prediction module for those young projects that will turn out unsuccessful at exit, most likely as a result of poor quality at entry. Lacking a quality at entry indicator, the most reasonable proxy for such an indicator are lags from approval to signing of greater than three months and from approval to effectiveness of greater than six months. However, the correlation of the outcome rating with those delays is not been well established and appears rather weak, so this third rule or module would not be permanent, meaning that after the projects have completed the first year, the rule should no longer apply to them.

The country breakdown of the three watch list sets is in Table 1 of Annex 2.

In view of the well-established correlation between the outcome and the CPIA and recent country record, the percentages of project failure implied by the latest candor gap figures (see Table 1 of Annex 2) were used to calculate a predicted MU- rate for each country. This was then compared to the watch list generated by the model. The two lists have a few significant differences (China and Vietnam in particular), but overall there is a good consistency with 106 of projects identified at the regional level compared with 112 corresponding to 38 percent of expected MU- outcome based on the candor gap figure, as of July 1, 2014.

The composition of the list, broken down by module components, is summarized in Table 6.

Table 6. FY15 Watch List Projects for EAP

Age group	**Three out of 6 flags but not last IP/DO MU-**	**Three flags and last IP/DO MU-**	**Only MU- IP/DO, but not three flags**	**Approval to signing/effect. lags**	**Total**
FY15 closings	13	13	4	N/A	**30**
< 1year old	0	0	2	9	**11**
Others	23	17	25	N/A	**65**
Total	**36**	**30**	**31**	**9**	**106**
%	34%	28%	29%	8%	**100%**

The breakdown of the watch list by module is as follows: 62 percent of the list is generated by the flag-based module, including 28 percent that are also rated MU- for IP and DO in their last ISR. Slightly less, 57 percent, would have been picked up by the problem project module alone. By age group, the projects in their closing years seem overrepresented compared with first-year projects (48 percent versus 19 percent), or even with the remaining group (37 percent), but several of these projects will not close and will be extended into the following year. It also seems sensible to be more cautious for the exit year than for the first-year projects.

The problem project, flag-based module and also a module of projects with delays to signing and effectiveness were used to create a watch list of EAP projects based on July 1, 2014 data. A review methodology was designed based on IEG ICR review criteria, i.e., relevance of objectives and design, efficacy, and efficiency. The reviews were prepared in batches and forwarded by the regional vice-president to the task team leaders and their practice managers. Once the projects on the watch list begin to exit, it will be possible to assess the accuracy of the prediction model when applied to an ongoing portfolio.

IV. Conclusion

The downgrading of projects at the closing from moderately satisfactory to moderately unsatisfactory has been a persistent problem in the World Bank and a particular problem in the EAP region since 2012. Through analysis of the projects that exited the EAP portfolio in FY12 and FY13, this paper has derived a prediction model that appears to improve on existing methods for assessing the downgrade risk.

The Bank has recently revised its project monitoring system and has introduced Standard Reports that provide management with information on project performance based on the flags self-reported by task teams in the ISRs. The problem project module is already used in the Standard Reports and, as this research has shown, continues to have value in identifying specific projects for attention in order to try to avert an unsatisfactory outcome.

The Standard Reports also recognize that a sizeable number of projects that ultimately end up with outcome ratings of unsatisfactory will largely coast through their implementation phase rated moderately satisfactory. There is a category of reports referred to as “Soft MS Rating” in the Standard Reports that identifies projects whose IP or DO rating is rated moderately satisfactory for more than 12 months. Research does not show any correlation between time rated moderately satisfactory and ultimate outcome. We therefore propose that the definition of the “Soft MS Rating” in the Standard Reports be replaced with the “flag-based module” developed here, which is more robust in predicting unsatisfactory outcomes. This will help to fill the gap in current portfolio monitoring and provide managers with a list of outlier projects that are at risk so that they may get timely attention to avert a moderately unsatisfactory or lower outcome rating.

Acronyms and Abbreviations

CPIA	Country Policy and Institutional Assessment
EAP	East Asia and Pacific region
FPD	Financial and private sector development
HDN	Human Development Network
ICR	Implementation Completion and Results report
IEG	Independent Evaluation Group
IP	Implementation progress
ISR	Implementation Status Report
M&E	Monitoring and evaluation
MNA	Middle East and North Africa region
MS	Moderately satisfactory
MS+	Moderately satisfactory or higher
MU	Moderately unsatisfactory
MU-	Moderately unsatisfactory or lower
PREM	Poverty Reduction and Economic Management (vice-presidency)
SDN	Social Development Network
TTL	Task team leader

Annex 1 Table 1

# Rated	Cambodia	China	EAP	Indonesia	Kiribati*	Lao PDR	Mongolia	Papua New Guinea	Philippines	Samoa*	Solomon Islands	Thailand	Timor-Leste	Tonga	Vietnam	Total Region
FY08-FY10 (# IEG Rated)	2	39	1	17		3	4	1	7	1		1	5	1	12	96
FY11-FY13 (# IEG Rated)	8	24	1	22	1	6	6	1	8	1	1	1	4	1	15	101
Grand Total (#)	**10**	**63**	**2**	**39**	**1**	**9**	**10**	**2**	**15**	**2**	**1**	**2**	**9**	**2**	**27**	**197**
FY08-FY10 (MU- %)	100%	8%	0%	41%		0%	0%	0%	57%	0%		0%	40%	0%	0%	21%
FY11-FY13 (MU- %)	25%	17%	0%	32%	100%	33%	33%	0%	75%	0%	0%	100%	100%	100%	27%	34%
Grand Total (MU-%)	**40%**	**11%**	**0%**	**36%**	**100%**	**22%**	**20%**	**0%**	**67%**	**0%**	**0%**	**50%**	**67%**	**50%**	**15%**	**27%**
Candor Gap Set #**	4	17	#N/A	19	1	5	4	#N/A	8	#N/A	#N/A	1	3	1	12	77
Candor Gap Set: % MU-	50.0	23.5	#N/A	36.8	100.0	20.0	25.0	#N/A	75.0	#N/A	#N/A	100.0	100.0	0.0	25.0	38
FY12-13 Sample #	6	12	1	11		4	5	1	6		1	1	4		10	62
FY12-13 Sample % MU-	33	33	0	64		50	40	0	83		0	100	100		30	48
Predicted MU- %	50%	24%	100%	37%	100%	20%	25%	100%	75%	100%	100%	100%	100%	0%	25%	38%
# IPF projects in portfolio	9	108	1	34	4	20	13	10	14	8	5	1	4	4	52	298
Sample Size for Watch list†	5	25	1	13	4	4	3	10	11	8	5	1	4	0	13	112

Notes: FY08-FY13 data are from IEG RAP report.

* Kiribati and Samoa were not in the FY12-FY13 set.

**The calculation of candor gap for this report used the number of projects reviewed in the 18-month period 07/01/2103-12/31/2014.

†Watch list sample size was calculated using the candor gap set.

Annex 2 Table 1

Breakdown of EAP FY15 Watch List by Country Management Unit

Country	Projects closing in FY15 (Set-1)	Project age < 1 year (Set-2)	Remaining Projects (Set-3)	Total	Forecast Size
Cambodia	2	1	1	4	5
China	12	2	20	34	25
EAP				0	1
Indonesia	4	2	10	16	13
Kiribati			2	2	4
Lao			5	5	4
Mongolia	1			1	3
Pacific Islands			1	1	
Papua New Guinea	1		3	4	10
Philippines	2		5	7	11
Samoa		2	1	3	8
Solomon Islands	1		1	2	5
Thailand				0	1
Timor-Leste	1		2	3	4
Tonga			1	1	
Vietnam	6	4	13	23	13
Watch list Total	**30**	**11**	**65**	**106**	**112***
Total Count	**63**	**59**	**175**	**297**	
%	**48%**	**19%**	**37%**	**36%**	

* Sum of the forecast is 106 but the target for the region is 112 because we took the percentage for the entire portfolio.

Annex 2 Table 2

Breakdown of projects in the Watch list	
# with MU- for IP/DO or at least 3 key flags rated MU- out of 6**	97
# with lag to effectiveness > 6 months or lag to signing > 3 months*	9
Total	**106**

Set	Definition	MU- IP/DO or ever has a MU- for at least 3 or the 6 Flags**	Add ons	Total
1	Projects closing in FY15	30		30
2	Projects less than one year (as of 07/01/2014)	2	9*	11
3	Remaining projects	65		65
Total		97	9	106

*Lag to approval >6 or lag to approval to signing > 3 months
**6 flags: project management, procurement, M&E, counterpart funding, safeguards, financial management

Set	Only 3 Flags	3 Flags and MU- IP/DO	Only MU- IP/DO	Approval Lag	Total
1	13	13	4	N/A	30
2	N/A	N/A	2	9	11
3	23	17	25	N/A	65
Total	36	30	31	9	106

Bibliography

Cevdet Denizer, Daniel Kaufmann, Aart Kraay, 2011, “Good Countries for Good Projects: Macro and Micro Correlates of World Bank Performance,” Policy Research Working Paper #5646, World Bank, Washington, DC.

Independent Evaluation Group (IEG), 2015, *Results and Performance of the World Bank Group 2014,* World Bank, Washington, DC

Jurgen Rene Blum, 2014, “What Factors Predict How Public Sector Projects Perform?: A Review of the World Bank’s Public Sector Management Portfolio,” Policy Research Working Paper #6798, World Bank, Washington, DC.

Patricia Geli, Aart Kraay, Hoveida Nobakht, 2014, “Predicting World Bank Project Outcome Ratings,” Policy Research Working Paper #7001, World Bank, Washington, DC.

Peter Moll, Patricia Geli, Pablo Saavedra, 2015, “Correlates of Success in World Bank Development Policy Lending.” Policy Research Working Paper #7181, World Bank, Washington, DC.

Quality Assurance Group (QAG), 2006, “Review of the Flag Risk System,” First Phase Draft Report, unpublished.