What influences World Bank project evaluations?[*]

Christopher Kilby
Villanova University

chkilby@yahoo.com

Katharina Michaelowa
University of Zurich & Jawaharlal Nehru Institute of Advanced Study, JNU
katja.michaelowa@pw.uzh.ch

May 13, 2016

Working Paper

Abstract: In July of 2011, the World Bank's Independent Evaluation Group (IEG) made the World Bank Project Performance Ratings database available to the public. With global coverage and spanning the range of sectors important for economic development, these data are a valuable resource for research about development effectiveness, what works and what does not. This paper first provides a brief overview of the IEG's rating process and the classifications used by the World Bank that may help scholars interested in evaluation, but unfamiliar with World Bank terminology, to use and interpret the database. Second, this paper examines whether geopolitical or institutional factors influence project performance assessments. We focus on how IEG selects projects for performance assessments and what factors influence project ratings in those assessments. We find some evidence that bureaucratic factors influence selection but only one geopolitical variable—UNSC non-permanent membership—influences IEG ratings.

JEL codes: F33, H43, O19, O22
Key words: World Bank; evaluation; geopolitics

[*]Thanks to Alex H. McKenzie and Martha Ainsworth of the World Bank Independent Evaluation Group for detailed information about the data and World Bank procedures.

1. Introduction

In July of 2011, the World Bank's Independent Evaluation Group (IEG) made its evaluation ratings—the IEG World Bank Project Performance Ratings database—publicly available as part of the World Bank's new "Access to Information" policy (World Bank 2012A). This database now includes outcome and other ratings on more than 8500 completed development projects and programs with funds administered by the World Bank, by far the largest set of development project evaluations available from a single organization. Although there has been some previous academic work using these data, the difficulties of obtaining official permission limited access by World Bank outsiders so that the bulk of earlier research has been done by insiders.[1] With the current open access policy, we expect to see a wave of new academic research.

Evaluation of project performance is to a degree a subjective enterprise and there is evidence of systematic bias in some World Bank assessments at project appraisal (Pohl *et al*. 1992) as well as changes over time in the evaluation system (Grasso *et al*. 2003). In addition, since the literature suggests a significant link between geopolitics and World Bank lending decisions (Dreher et al. 2009; Kilby 2013A), a similar influence on ratings appears plausible.

Evaluation bias or process changes are of particular concern to researchers if these are correlated with the researchers' key explanatory variables. A few researchers have devised clever ways to check for evaluation bias related to their particular question (Dreher *et al*. 2013) but it is

[1] Outsider papers includes Buntaine and Parks (2013), Dreher *et al*. (2013), Girod and Tobin (2011), Kilby (2015), Malik and Stone (2016), Michaelowa and Borrmann (2006), Sud and Olmstead-Rumsey (2012), and Winters (2014). Insider papers include Blum (2014), Bulman *et al*. (2015), Chauvet *et al*. (2010), Chauvet *et al*. (2015), Cruz and Keefer (2013), Deininger *et al*. (1998), Denizer *et al*. (2013), Dollar and Levin (2005), Dollar and Svensson (2000), Geli *et al*. (2014), Guillaumont and Laajaj (2006), Isham and Kaufmann (1999), Isham *et al*. (1997), Kaufmann and Wang (1995), Limodio (2011), Malesa and Silarszky (2005), Moll *et al*. (2015), Pohl *et al*. (1992), Ralston (2014) and Smets *et al*. (2013).

an issue that anyone using these data must address if they wish to draw meaningful conclusions from their analysis.

For many World Bank projects, the database contains more than one set of ratings. The Implementation Completion Report (ICR)—written by or under the direction of the Team Task Leader (TTL) in charge of supervising project implementation in its final stages—includes one set of ratings. IEG selects some projects for additional review and, in these cases, IEG staff and consultants prepare a Project Performance Assessment Report (PPAR) that includes a second set of ratings. This paper explores IEG's role in evaluation, examining both the PPAR selection process and how PPAR ratings differ from ICR ratings. In addition to considering the effect of project and country characteristics, we examine the role of internal institutional forces and external geopolitical factors that may reflect donor influence. In doing so we follow a three-fold objective: First, we intend to contribute to the understanding of the World Bank's evaluation procedures; second, we wish to provide an overview of existing evaluation biases; and third, we offer a choice of potential instruments that may be useful for researchers using the World Bank evaluation data in the future.

The remainder of this paper is structured as follows. Section 2 provides an overview of World Bank project evaluation. Section 3 analyses PPAR selection in a hazard model framework. Section 4 explores the determinants of PPAR ratings. Section 5 concludes.

2. Project Evaluation at the World Bank

Evaluation is the final step in the World Bank's project cycle. It can serve two functions. First, it allows lessons learned to shape country assistance strategies and the identification, preparation, and implementation of new investment projects and program loans. Second, it

promotes accountability, holding project staff accountable to management and, especially, holding management accountable to stakeholders. World Bank President Robert S. McNamara formalized ex post project evaluation at the World Bank (initially referred to as an audit) with the creation of the Operations Evaluation Unit in 1970. This unit was elevated to a separate division in 1971 and then to a department—the Operations Evaluation Department (OED)—in 1973. Under pressure from the U.S. Congress and its General Accounting Office and with the support of McNamara, in 1975 OED top management switched from a Vice President appointed by and reporting to the President to Director General, Operations Evaluation (DGO) appointed by and reporting to the World Bank Board of Executive Directors (EDs). To safeguard the independence of the evaluation department and the content of its reports, the DGO can only be removed by the Board and cannot return to the Operations side of the World Bank. Operations staff and management (and borrowing governments) can comment on but not directly edit draft reports; after incorporating comments "as appropriate," IEG attaches borrower comments to its final version (IEG 2015A, vi). In November of 2005, OED merged the much smaller evaluation units for World Bank's International Finance Corporation and Multilateral Investment Guarantee Association and was reincarnated as IEG (IEG 2015B).[2]

As part of this move toward systematic evaluation, starting in 1973 the World Bank has required a Project Completion Report (PCR)—since 1995 more accurately called an ICR—for all World Bank-funded projects. Written by or under the direction of the TTL in charge of project

[2] For a more detailed history of OED, see Grasso *et al*. (2003). For simplicity, in the text below we use IEG to refer to the evaluation department even before 2005.

supervision at the final stages of implementation, PCRs/ICRs now include categorical project ratings.[3]

One of the key functions of IEG is to prepare PPARs, ex post evaluations of World Bank-funded projects. PPARs start with a desk review of the ICR and project files and typically include field research in the borrowing country. Figure 1 illustrates this evaluation process within the overall project timeline. IEG's target PPAR rate was initially 100% (essentially achieved by 1981 (IEG 2015B, 2)) but selective ex post evaluation was introduced in 1983 and the target rate was gradually reduced over time to 50%, then 40% and finally to 25% in 1997 (Grasso *et al*. 2003, 178).[4] PPARs also include performance ratings that can differ from ICR ratings and do so 16% of the time in our data.[5] Whether or not IEG conducts a PPAR, it records performance ratings for virtually all completed World Bank-funded projects in its database (see IEG 2015B).

[Figure 1 about here]

The availability of ratings in the IEG database merits some discussion. Up through the end of 1996, ICR ratings appear in the database as Project Completion Ratings (PCRs). Phased in starting in early 1995, IEG policy shifted to include a "validation" step before ratings—termed Evaluation Memorandum (EVM) ratings and later Evaluation Summary (ES) ratings—enter the database (Denizer *et al*. 2013). If IEG conducts a PPAR, it includes the associated rating in the

[3] Early PCRs were joint products of Operations and IEG staff and did not include ratings; IEG has since imputed ratings from the written reports (IEG 2015B, 2).

[4] This reduction was driven by an expanding World Bank project portfolio but also increasing demands on IEG resources to generate other evaluation products at the country and sector levels (Grasso *et al*. 2003). More than 70% of IEG's budget was devoted to PPARs in 1976 (Grasso *et al*. 2003, 169); the figure was less than 10% of IEG's USD 34 million budget in 2011 when the time devoted to a PPAR averaged six staff weeks (IEG 2011, 38). The PPAR rate has differed across activities depending on their "novelty" and importance. For example, IEG kept the PPAR rate on the then-new structural adjustment programs at 100% in the 1980s even as it reduced the PPAR rate on investment projects.

[5] Calculated from 1590 IBRD/IDA-funded projects with overall outcome ratings measured as "Satisfactory" or "Unsatisfactory." For our slightly smaller estimation sample, the figure is 17.6%.

database. For simplicity, we refer to this as a PPAR rating though the IEG database uses the code PAR based on an older name for these reports (Project Assessment Report).[6] In early years, the database reports only PPARs systematically replacing previous PCR ratings. Starting in FY 1983, some projects report an initial PCR rating as well as a PAR and from 1995 onward virtually all projects report an initial PCR/EVM/ES rating.[7] Figure 2 shows the type of the first ratings reported in the database (in % of total first ratings).

[Figure 2 about here]

According to IEG, the selection of projects for the in-depth PPARs is non-random (Grasso *et al*. 2003). PPAR sample selection reportedly depends on several factors including particularly good or bad outcomes, sectors subject to IEG review, and geography.[8] IEG staff typically complete PPARs 1 to 5 years after the project closes (i.e., the close of disbursement of the IBRD loan or IDA credit).

Ratings have expanded over time from a single 0/1 outcome rating into numerous categorical ratings. Most research and policy discussion continues to focus on the original

[6] PPAR is the report (also referred to as a Project Performance Audit Report or just Project Audit Report in the earlier IEG literature) while PAR is the database name for its associated ratings.

[7] In some cases, a second PPAR rating is also available, either because of a transition in rating systems right around the time IEG completed the PPAR or as the result of a second PPAR to reassess long run performance some years later. These details are relevant for discussions below. To allow for these details and possible data revisions, we take the project's last PCR/EVM/ES entry as its ICR rating and the project's first PPAR entry as its PPAR rating. In all but a handful of cases, these are also the project's only ICR and PPAR entries.

[8] Despite some selection based on ICR ratings, the overall share of projects initially rated satisfactory does not vary too much between those projects that get reevaluated and those that do not: ICR ratings average 72% satisfactory for projects with no subsequent PPAR versus 78% for projects with a subsequent PPAR. Starting in the mid-1980s, IEG began "group audits" for sequences of projects in the same sector in a given country and "cluster audits" of similar projects in several neighboring countries (Grasso *et al*. 2003, 48-49). The selection of projects for a PPAR is ultimately the decision of IEG division chiefs but with input from IEG staff (Grasso *et al*. 2003, 49).

outcome rating reduced to a binary variable. Studies examining ratings in both raw and binary forms (Denizer *et al.* 2013, Dreher *et al.* 2013) generally do not find compelling reasons to use the more fine-grained version except Sud and Olmstead-Rumsey (2012) who argue that rating system changes have facilitated illusory improvements in performance. The outcome rating ostensibly measures project outcomes relative to objectives stated in the project appraisal and loan documents though there is evidence that an economic rate of return cut-off of 10% (i.e., an absolute standard) is used to distinguish between "Satisfactory" and "Not Satisfactory" where such figures are available (Kilby 2000; Grasso *et al.* 2003, 11).[9]

IEG ratings cover a number of topics in addition to the overall project performance ("outcome") rating. Table 1 provides details on these ratings.

[Table 1 about here]

3. PPAR Selection

As noted above, the selection of projects for performance reviews depends on a number of factors and is not simply random. Since IEG periodically reviews a long sequence of projects, a project might be selected for a PPAR years after completion; the longest interval in the database between ICR and PPAR is almost 18 years.[10] Of the 158 countries and territories with projects included in the IEG database, fourteen have not yet had a project with a completed PPAR.[11] In

[9] This pattern is apparent in IEG (2010); Appendix B reports that only 12% of projects with an economic rate of return above 10% were rating "moderately unsatisfactory" or lower. Denizer *et al.* (2013) also argue that World Bank procedures promote applying relatively uniform standards to goal setting and evaluation.

[10] The Vocational Training and Technological Development Project in Uruguay was approved in 1978 and completed in 1986. Its Project Completion Report (equivalent to an ICR) was issued in 1988 and the project was included with two subsequent Uruguayan education projects in a 2006 PPAR. The longest interval in our estimation sample is 12.75 years.

[11] Countries with ICRs (number of ICRs in parentheses) but no PPARs as of September 30, 2013, are: Angola (15), Cape Verde (22), Sao Tome and Principe (13), Tonga (5), the Bahamas (5),

principle, a project always has some positive PPAR selection probability, no matter when it was completed. For this reason, a hazard model that treats unselected projects as not-yet-selected is appropriate. We start with a baseline model that includes the variables suggested as factors driving selection by IEG itself (such as ICR rating), IEG's assessment of the quality of this initial rating, and some basic project characteristics such as project size, geographical region, project type, etc. Moreover, we include additional variables reflecting institutional peculiarities and bureaucratic preferences that may be relevant for the selection of a project for reevaluation. The latter notably includes the attractiveness of the country in which the project is located (proxied by tourist arrivals) and a dummy for the ICR being completed in the month of June, a month in which these numbers tend to be high as this is the last month of the World Bank's financial year. We expect that given the high numbers completed simultaneously, only a smaller share of these projects may be reevaluated. We also include some recipient-country characteristics to account for the general economic and political context. All variables will be explained in more detail in the data section below.

However, the baseline model excludes all variables reflecting geopolitical importance and the interests of powerful shareholder countries. We consider these variables separately thereafter, as they are of special interest given the strong evidence for the role of these factors on other areas of World Bank activity.

To start with, we examine the role of some basic project characteristics later used in the multivariate hazard model using graphs of cumulative hazard rate functions (presenting simple

Grenada (9), St. Kitts and Nevis (5), St. Vincent and the Grenadines (6), Turkmenistan (3), West Bank and Gaza (40), Kosovo (20), Namibia (2), Montenegro (5), and South Sudan (1).

bivariate relationships).[12] Figure 3, Panel A plots separate cumulative probabilities of a PPAR for projects with IDA funding (including "blend projects" with both IBRD loans and IDA credits) and projects with no IDA funding. During the first few years after an ICR, non-IDA projects are slightly more likely to have been selected but past the 6-year mark, IDA projects are more likely to have been selected. Panel B presents the graph for investment projects and program loans, illustrating the much higher frequency of PPARs for the latter group. Panel C confirms that PPARs are more likely for projects initially rated "Satisfactory" ("Moderately Satisfactory" or better in detailed ratings) than for those initially rated as "Unsatisfactory" ("Moderately Unsatisfactory" or worse in detailed ratings). Panel D reveals that projects where the ICR was rated as low quality by IEG in its initial validation step (6% of the sample) are much more likely to be subjected to additional review. Panel E highlights variation in PPAR hazard rates by the number of tourist arrivals per year in the project country. The odds that IEG staff will complete a PPAR (which typically includes travel to the country) rise faster for projects in countries that are popular with tourists. Finally, Panel F illustrates differences between regions, with the lowest PPAR hazard rate for projects in Latin America and the Caribbean (LAC) and the highest rate for projects in South Asia (SA). Of course, all the above patterns could reflect compositional differences so we now turn to estimation results.

[Figure 3 about here]

Data description

[12] We include graphs for variables that we think will interest researchers and World Bank management.

Along with Figure 3, Table 2 and Figures 4-5 describe the data used in the PPAR selection analysis. Table 2 presents descriptive statistics for all variables. For the baseline specification that excludes geopolitical variables, the sample includes 5155 projects. Project approval dates run from 1963 to 2011 and ICR dates from 1983 to 2012.[13]

[Table 2 about here]

The sample above is determined in part by the availability of ICR ratings. To allow for the possibility of an ICR followed by a PPAR, we limit the sample to cases where the PPAR is not the first recorded evaluation rating. As shown in Figure 2 above, the first ICR rating (a PCR, EVM, or ES) appears for a project evaluated in FY1983 (calendar year 1982); in earlier years, the PPAR rate was 100% and the IEG database only reports PPAR ratings. Only by FY1994, virtually all projects report an ICR rating. We further limit the baseline sample to observations with inflation rates below 1000% to insure the robustness of inflation results discussed below.

The last 7 variables in Table 2 reflect geopolitical importance or the interest of powerful shareholder countries. As mentioned above, these variables are not part of the baseline specification; we introduce and discuss separately each of them as candidates for the potential political and institutional biases in which we are interested. The number of observations is smaller for some of these variables due to limited data availability (by country or year).

The analysis examines the (censored) time between ICR and PPAR.[14] Figure 4 depicts the distribution of projects by years between ICR and PPAR for the 1590 projects in the estimation

[13] In a few cases, there is a simplified ICR prior to the final ICR or a second PPAR (referred to as a PPRR – Project Performance Reassessment Report). We always use the last ICR and the first PPAR in our analysis.

[14] The hazard model sample includes all years with available data whereas the PPAR rating sample below excludes FY2002 (which proves to be an outlier in that setting). Hazard model results are largely unchanged if we also exclude FY2002 here.

sample with PPARs. This ranges from 5 days to 12 years, 9 months with a mean of 2½ years (median 2 years). Figure 5 provides similar information for the 3784 projects for which IEG has yet to conduct a PPAR (as of September 30, 2013). Time since ICR ranges from 276 days to 24 years, 7 months with a mean of 12 years (median 11 years, 10 months). As the graphs show, the risk of a PPAR rises and then tails off exponentially over time while the group of projects not yet selected for a PPAR shows no particular pattern in terms of time since ICR (other than the decline in the PPAR selection rate over the past thirty years).

[Figures 4-5 about here]

The factors that we consider as potential determinants of PPAR selection have already been mentioned above, but require some more detailed explanations. Section 2 suggests that an important factor should be the initial rating. In this context, we focus primarily on a dichotomous version of the overall ICR rating of project performance, *Outcome* (*ICR*), which indicates a 73% success rate in this sample. However, we will also consider a six-point rating scale that was introduced in 1995 (see Column 4 of Table 3); in that (smaller) sample, 5% are "Highly Satisfactory," 44% are "Satisfactory," 25% are "Moderately Satisfactory," 9% are "Moderately Unsatisfactory," 15% are "Unsatisfactory," and 2% are "Highly Unsatisfactory."

In addition to the ICR rating, a number of other institutional/bureaucratic variables may play a role in PPAR selection. The *IDA* dummy equals 1 for slightly more than half the projects; we include this even though Figure 3 shows only small differences between IDA- and IBRD-funded projects. Conversely, Figure 3 does suggest that *ICR quality* impacts PPAR selection. For 94% of projects in the sample, this variable indicates high quality ICRs; to avoid loss of observations, this includes imputed values if the IEG database does not report an ICR quality rating (see note A in Table 2 for details). We include the amount the borrower owes to the World Bank

(*log World Bank debt* in billions of 2005 USD) in case the World Bank treats major borrowers differently; calculated as a three year moving average (from two years before evaluation to evaluation), the variable averages 0.68 (USD 2 billion) and ranges from -4.75 (USD 8.7 million for Grenada in 1990-1992) to 3.53 (USD 34 billion for India in 1994-1996). The variable *log Project size* (in millions of 2005 USD) may be relevant if, for example, larger projects are more likely to be selected for PPARs because more resources were invested, making the project more important to both bank and borrower. The average is 4 (USD 55 million) and the range is from -0.63 (USD 0.5 million for a project approved in FY2000 for Tajikistan) to 8 (USD 3 billion for a single tranche program loan approved and disbursed in December of 1997 for Korea at the peak of the Asian Financial Crisis).

Three other bureaucratic factors may play a role in the selection of projects for PPARs. First, we include the number of ICRs in the same country in a given fiscal year to allow for IEG's cost-saving practice of "cluster audits"; we take the log because of a skewed distribution driven by large borrowers. The resulting variable *log # WB projects* ranges from 0 (the project itself) to 2.89 (18 ICRs in 1997 and 1998 in India). Second, we account for patterns across the World Bank's fiscal year with a June ICR dummy. Eighteen percent of ICR ratings appear in the IEG database in June, the last month of the World Bank's fiscal year. Third, PPARs typically include a trip to the borrowing country (a "mission" in World Bank parlance). IEG staff have input into the PPAR selection decision so we also include the variable *Tourism*, the log of the annual number of tourism arrives in the country. This variable correlates with the desirability of the country as a travel destination. *Tourism* ranges from 9.53 (an annual average of 10,000 tourist arrivals in São Tomé and Principe) to 17.51 (an annual average of 4 billion tourist arrivals in China) with a mean of 13.98 (1.2 million arrivals). Given the skewed distribution of tourism arrivals (particularly with

China in the sample), taking logs is advisable. Results reported below are not sensitive to excluding China (and/or India) from the estimation sample.[15]

The next set of variables captures characteristics of the borrowing government and country. *Years in office* is the number of years the executive has been in office at the time of the first project evaluation and ranges from 1 to 45 (King Hussein of Jordan in 1998). *Freedom House* is the three year moving average of the 1-to-7 Political Rights and the 1-to-7 Civil Liberties indicators, where lower values indicate fewer restrictions, i.e., a more democratic regime.

We measure country size with *log Population*, the log of a three-year moving average of population. This averages 17.1 (roughly 27 million people) and ranges from 11.48 (Grenada) to 21.02 (China). The level of development is captured with *log GDP PC*, the log of a three-year moving average of GDP per capita measured in year 2000 USD. This averages 7 (just below USD 1100 per capita) and ranges from 4.78 (USD 119 per capita in Ethiopia in 1993-1995) to 9.31 (USD 11,075 per capita in Hungary in 2008-2010). We also include *GDP growth*, the three-year moving average of GDP growth, with a mean of 4.6% and a range from -17.49% (Rwanda 1993-1995) to +28.65% (Azerbaijan 2006-2008). *Inflation* is based on the GDP deflator (again, a three year moving average), with a mean of 0.16 (16%) and a range from -0.13 (-13% in Ecuador in 1999-2001) to 8.07 (807% in Democratic Republic of the Congo in 1990-1992). As noted above, we exclude some hyperinflation cases (*Inflation*>1000%) from the sample as results can be very sensitive to these.[16]

[15] Tourism data are sparse so we take the average across available years for each country. Thus, this is a purely cross-sectional variable. The specification also includes log population so that the tourism coefficient would be the same if we measured the desirability of a destination instead as tourist arrivals per capita.

[16] For example, in the unrestricted sample the estimated coefficient on an inflation interaction switches from positive and significant to negative and significant when dropping just one high-inflation observation.

The baseline variables include a final set of dummies. *Program Loan* equals 1 if the "project" is a program loan (a Structural Adjustment Loan or a Development Policy Loan) which account for 14% of the sample. *SIL* is equal to 1 for investment projects financed via a Specific Investment Loan, one of the major types of investment loans. We include a set of regional dummies (with Sub-Saharan Africa as the omitted category in the specifications estimated). Regional effects can be driven by real world differences or functional differences between World Bank administrative divisions.

The final seven geopolitical variables in Table 2 that are not part of the baseline specification cover aid flows, alignment measures, and memberships on powerful committees. We measure aid flows either as the log of one plus U.S. bilateral total official gross disbursements in millions of 2005 USD (*log US Aid*—which should reflect U.S. interests) or the log of one plus G7 bilateral total official gross disbursements in millions of 2005 USD (*log G7 Aid*—which should capture broader G7 interests).[17] The U.S. variable averages 4.07 (USD 58 million) and ranges from no aid to 8.58 (USD 5.3 billion to Egypt, average from 1989 to 1991). The G7 variable averages 6.17 (USD 478 million) and ranges from 0.98 (USD 2.7 million to Afghanistan, average from 1984 to 1986) to 8.93 (USD 7.5 billion to Egypt, average from 1989 to 1991). *US Military Aid* is a dummy variable, equal to 1 if U.S. military aid exceeds 2 million 2005 USD which is the case in just over 40% of the observations. We use a dummy variable because military aid has both a skewed distribution (which argues for logs) and a large number of zeros (which argues against logs). The USD 2 million threshold captures non-trivial military aid flows though results are not sensitive to varying the threshold.

[17] We add one before taking logs to avoid log of zero; we also disregard the few negative values reported by the OECD since this is defined a gross measure.

UN voting alignment measures draw on votes identified as important to U.S. interests in the annual report by the U.S. State Department to the U.S. Congress. For any two countries, we code matching votes as 1, opposite votes as 0, and cases where one party abstained or was absent but the other party did vote as 0.5. Alignment is the average of these values across all relevant votes in a given year but matched to other data in the next year since UN votes fall in the last quarter of the calendar year, often in November and December. For the G7, the alignment measure is an average of the alignments with each of the seven countries.

The final two variables reflect membership in important committees. *UNSC* (*ICR*) is equal to 1 if the country holds a non-permanent United Nations Security Council seat in the year of the ICR evaluation. The last variable is *WBEB* (*ICR*), a dummy variable indicating if the country held an Executive Director (ED) seat on the World Bank's Executive Board in the ICR year or either of the two previous years. With a 25 member board, many countries share an ED seat and their representative comes from one of the countries in the group. As Kaja and Werker (2010), Morrison (2013), and Kilby (2013B) demonstrate, this can be a powerful position that improves access to World Bank resources for the ED's home country. Any of these characteristics might play a role in the selection of projects for performance assessment by IEG.

Hazard model estimation

Table 3 reports hazard ratios from a parametric survival time model using a proportional hazard specification and the Weibull distribution as well as odds ratios from a logit selection model for comparison. Reported z-statistics reflect standard errors clustered at the country level. Results

are similar using an accelerated failure-time model, alternative distributions, or a Cox proportional hazard model.[18]

Column 1 of Table 3 is the baseline hazard estimation covering the full available sample. Hazard ratios greater than 1 indicate a higher PPAR hazard rate (greater risk of a PPAR) while ratios below one indicate a lower PPAR hazard rate (lower risk of a PPAR).[19] All specifications include unreported fiscal year dummies based on the ICR date.

[Table 3 about here]

As foreshadowed by Figure 3, the risk of a PPAR is significantly higher for projects with a positive ICR outcome rating (*Outcome* (*ICR*)). This could reflect a desire to learn lessons from particularly successful projects, to correct overly optimistic ICR ratings, or to incentivize TTLs to report accurate ICR ratings. The source of funding (IBRD versus IDA) does not influence IEG's decision to conduct a performance assessment—once we control for income level.[20] Parallel to the estimated hazard ratio for *Outcome* (*ICR*), low *ICR quality* is a strong predictor of a PPAR, presumably because IEG staff feel that the learning phase of the project cycle is short circuited in the case of a low quality Implementation Completion Report (or, again, to incentivize TTLs).

[18] Results are comparable for the Gompertz, loglogistic, and lognormal distributions. A likelihood ratio test rejects the shape assumption of the exponential model in favor of the more general Weibull though results are again similar. Estimates are similar with the gamma distribution but standard errors are larger and a few variables (notably *June ICR*) fail to reach statistical significance.

[19] Hazard rates were first introduced in a medical context where the event in question was death. This is why the technical language speaks of 'survival models' and of 'risk,' no matter whether the event in question (here a PPAR) is desirable or not.

[20] Consistent with the hazard ratio below unity for GDP per capita in this specification (discussed below), if we omit GDP per capita the estimated *IDA* hazard ratio becomes significantly greater than one. So IDA projects—to low income countries that qualify for IDA's concessional credits—do have a greater PPAR risk but this is part of the more general pattern that PPAR risk declines as income increases.

The hazard ratio for *log World Bank debt* is greater than one though only marginally so and becomes insignificant in the slightly smaller samples of Columns (4) and (5). Thus, there is only weak evidence that World Bank exposure in a given country influences the decision to conduct a PPAR. As anticipated, larger projects do have a higher PPAR risk; the hazard ratio for *log Project size* is significantly greater than one at the 1% level. Consistent with IEG's practice of clustering PPARs by country to control costs, the hazard ratio for *log # WB projects* is significantly greater than one, indicating a higher incidence of project performance assessment reports when a country has several projects completed in the same year.

The flood of ICRs at the end of the World Bank's fiscal year does appear to reduce PPAR selection risk ("safety in numbers"). The hazard ratio for *June ICR* is significantly and substantially below one. To get a sense of the relative magnitude of this effect, we can compare it to a few other results. The estimated hazard ratio for *June ICR* indicates a 15% decrease in the PPAR hazard rate for projects with ICRs completed in June relative to ICRs completed in other months, *ceteris paribus*. The estimated hazard ratio for *Outcome* (*ICR*) indicates a 48% increase in the PPAR hazard rate for projects with satisfactory ICR ratings as compared to those with unsatisfactory ICR ratings, *ceteris paribus*. The estimated hazard ratio for projects with low quality ICRs (discussed below) indicates a 50% decrease in the PPAR hazard rate for projects with high quality ICRs as compared to those with low quality ICRs, *ceteris paribus*. Finally, the estimated hazard ratio for projects financed via Specific Investment Loans (SILs--again, discussed below) indicates a 17% increase in the PPAR hazard rate for projects financed via SILs as compared to those financed via other types of investment loans, *ceteris paribus*. Thus ignoring the direction of the effect, the magnitude of the impact of a June ICR is about one third that of a satisfactory *Outcome* (*ICR*) rating or low quality ICR but on par with being financed through a

SIL. One possible explanation is that IEG flags projects for PPARs when it receives the ICRs from Operations but it is reluctant to commit to too many PPARs in a given month. Alternatively, projects with ICRs completed at the end of the fiscal year might differ from other projects in some dimension not captured by our other covariates.

Our last institutional variable, *Tourism*, is also significant at the one percent level with a hazard ratio greater than one. *Ceteris paribus*, countries that attract tourists also attract PPARs—and the IEG staff and consultants who travel on mission to conduct these PPARs.

We turn now to variables that describe the borrowing country and its government. The hazard ratio for *Years in office* is significantly greater than one indicating a higher PPAR selection risk for more established governments.[21] The hazard ratio for *Freedom House* is significantly below one, indicating that projects in less democratic countries face a lower PPAR selection risk. Why do we see this pattern, i.e., a higher PPAR selection risk for projects under more experienced, more democratic governments? Note that this is not simply an effect driven by China with its large size and large number of projects (anyway, population size and project numbers are controlled for). One interpretation is that more experienced governments in more democratic countries complete their review of and comments on PPARs faster than governments that have been in office for less time or that are more autocratic. If so, the observed differences would have little to do with World Bank behavior. That is, these characteristics might impact how quickly the

[21] The impact of a single year is trivial but recall that years in office range from 1 to 45 so that a sizeable effect is predicted in some cases. Because years in office might be related to conflict (particularly very few or very many years in office), we also explored the impact of a conflict dummy derived from PRIO data at the ICR date. When included in this specification, the hazard ratio for conflict was marginally significant and less than one (conflict reduces the hazard rate for a PPAR). The magnitude of the hazard ratio for years in office is not impacted although it is now only marginally significant. Thus, the results reported above do not appear to be driven by omitted information about conflicts.

World Bank releases a PPAR but not whether or not a PPAR is undertaken in the first place. According to this story, the observed delay between the ICR and the PPAR—when a PPAR is observed—should depend on these characteristics but the odds of a PPAR being undertaken should not. Because of the censoring issue that motives the hazard model approach in this setting, only a rough test of this conjecture is possible. We return to this issue shortly when discussing Columns 2 and 3.

We also consider a number of other country characteristics. Projects in more populous countries have a lower PPAR hazard rate, *ceteris paribus*. The level of development (measured by *log GDP PC*) is a significant factor, with a lower PPAR risk in more advanced (richer) countries. *GDP growth*, on the other hand, does not appear to be a significant factor. Finally for inflation, we differentiate between investment projects and program loans since macroeconomic performance is a core focus of the latter. The estimated equation includes *Inflation*, *Program Loan*, and the interaction term *Program Loan*×*Inflation*. As Table 3 shows, inflation is not a significant criterion for PPAR selection in the case of investment projects but is for program loans: high inflation in the case of a program loan is associated with higher PPAR selection risk though only in the full sample (Columns 1-3). Finally consistent with Figure 3, the PPAR risk is higher in general for program loans than for investment projects.

We considered a number of other categorical project characteristics (type of financial instrument and sector) but only one proved relevant. The hazard ratio is significantly higher for SILs as compared to other types of investment loans in the full sample. Differences among the other types of investment loans were small and not statistically significant. Likewise, there were no significantly different hazard rates across sectors (*ceteris paribus*). To streamline the presentation, we omit all these factors other than *SIL*.

Looking at regional differences, we find that, *ceteris paribus*, PPAR selection risk is higher for projects in the *Europe & Central Asia* region relative to Sub-Saharan Africa. This result underscores the importance of controlling for compositional effects since the unconditional hazard rate for *South Asia* in Figure 3—not *Europe & Central Asia*—was the highest of all the regions.

As discussed above, it is possible to do a rough test of what drives hazard rate results—which projects are (eventually) selected or the length of the delay between the ICR and the PPAR for projects that are selected. The selection estimation suffers from right censoring and must be interpreted in this context (e.g., as the probability of being selected for a PPAR by September 30, 2013) but may still be informative. To this end, Column 2 of Table 3 presents odds ratios from a logit model of whether or not there is a PPAR by September 30, 2013 while Column 3 presents results from a hazard model estimated with only the uncensored data (i.e., the cases where we have a PPAR—see also Table 5 below). Column 2 should mirror effects in Column 1 that largely operate through selection while Column 3 should mirror those largely due to delayed (or accelerated) release of a PPAR. Comparing the three columns, the high hazard ratio for *Outcome* (*ICR*) in Column 2 coupled with the low (though insignificant) ratio in Column 3 suggest that the ICR rating influences PPAR selection—making the PPAR more likely—but not the length of time to complete a PPAR for selected projects. The low hazard ratio on the *IDA* dummy in Column 3 suggests PPARs for these projects take longer to complete. That *ICR quality* is significant only for the first two columns suggests it is a factor in selection only. The marginally significant effect of World Bank debt also appears driven by selection, a pattern seen again for project size, the number of World Bank projects, and June ICRs. In contrast, *Tourism* enters with odds/hazard ratios significantly greater than one in all three columns, indicating that projects in locations that

attract more tourists are both more likely to be selected for a PPAR and, once selected, more likely to have their PPAR completed quickly.

Looking at regional patterns, we find that PPARs are carried out more frequently for projects in Europe & Central Asia. No such effects are discernible for any other regions. However, in East Asia-Pacific as well as in the Middle East & North Africa seem to take more time. This may reflect some inefficiency in handling the evaluation process in these countries.

With respect to the higher PPAR risk for projects under more experienced, more democratic governments, the estimates do not support the "delay" story suggested above. If the effects were due to recipient government delays, the uncensored hazard model (Column 3) should drive the *Years in office* and *Freedom House* results while these variables should be insignificant in the selection logit (Column 2). However, we find just the opposite: Neither *Years in office* nor *Freedom House* is statistically significant in the uncensored hazard model while both are strongly statistically significant in the logit model. Apparently, the World Bank considers a government's experience and the level of democracy when deciding whether or not to undertake a Project Performance Assessment Review, preferring to conduct a PPAR when the government is more established and more democratic. Possibly, in these cases, PPARs are more agreeable to all partners because they expect better evaluation results. An alternative explanation could be that IEG expects a smoother evaluation procedure and a more fruitful learning effect for both the World Bank and the recipient country government from the results of the evaluations in such stable and democratic countries.

The last two columns of Table 3 look at the role of detailed ICR outcome ratings in PPAR hazard. Does the ICR outcome rating enter PPAR selection in a monotonic fashion (with the hazard rate declining consistently as the rating declines) or non-monotonically? One might

imagine that IEG targets projects: 1) with the most inflated ICR outcome ratings ("Highly Satisfactory"); 2) that were exceptionally effective so that lessons can be learned from success ("Highly Satisfactory"); and 3) that were exceptionally awful so that lessons can be learned from failure ("Highly Unsatisfactory"). To explore this issue, Column 4 reports results for projects with post-1994 ICRs which include a 6-point rating scale for the overall project performance variable *Outcome* (*ICR*). We include each rating category as a separate dummy variable; the omitted category is "Highly Unsatisfactory."

The hazard ratios support the first two conjectures: ICR outcome ratings appear to enter PPAR selection monotonically. PPAR risk is lowest for projects rated "Highly Unsatisfactory" and increases steadily. This suggests that IEG's strategy is to learn what to do, rather than what not to do.

Column 5 checks whether it is the reduced sample or splitting the outcome rating into subcategories that drives the few differences between Columns 1 and 4. Comparing Columns 4 and 5, it is clear that differences are sample-driven; Column 4 (with 6 rating categories) is nearly identical to Column 5 (2 rating categories but the same sample as Column 4). Higher inflation is no longer a significant factor in program loan PPAR risk (perhaps because of less variation in inflation since 1995) though the estimated hazard ratios are quite similar. The difference between investment projects financed via Specific Investment Loans (SILs) and other investment projects becomes insignificant in the later sample (although SILs still constitute about half of the data points).

The analysis thus far has only examined two of ten IEG project ratings (see Table 1 for a list and description of all ratings). After controlling for the ICR outcome ratings (including the five dummies discussed above) and the ICR quality rating, other ratings do not have a significant

impact on PPAR hazard. The one exception is the Borrower Compliance rating; the middle ratings ("Satisfactory" and "Unsatisfactory" up to 2006 plus "Moderately Satisfactory" and "Moderately Unsatisfactory" from 2006 on) enter the hazard function with positive and significance coefficient estimates as compared to the lowest category ("Highly Unsatisfactory"). One possible explanation could be that IEG expects the difficulty of the cooperation to extend to the evaluation, so that the implementation of a PPAR is considered less fruitful.

Estimated coefficients are not significantly different for program loans as compared to investment projects so compliance ratings appear to have the same impact on PPAR risk in adjustment and investment operations. Given that results remain substantially unchanged, we do not present the equations with the additional outcome ratings here.

Table 4 turns to a different issue, whether the interests of powerful World Bank shareholders play a role in PPAR selection risk. The table summarizes results from hazard ratio estimations that add, one at a time, political economy variables to the baseline specification from Column 1 of Table 3. Note that sample size differs by variable due to data availability. Each variable reflects values the year of the ICR outcome rating except for World Bank Executive Board representation which we have over three years as indicated in Table 2.

The first political economy variable is *log US Aid*, the log of total official gross bilateral aid disbursements from the U.S. The estimated hazard ratio is statistically significant but, perhaps surprisingly, greater than one.[22] The boarder G7 measure (*log G7 Aid*) also enters with an estimated hazard ratio greater than one. These results indicate a higher (rather than lower) PPAR

[22] We add one to all observations before taking logs (to avoid log of zero). A few large DAC values follow exceptional and questionable definitions (e.g., Egypt 1991 and Panama 1999 for the U.S.) so taking the log may be an important step to reduce the impact of outliers. In fact, without logs the estimated coefficient for *US Aid* is negative (though not significant in the full sample).

risk for countries that receive more U.S. or G7 bilateral aid. We explore this result further by including instead bilateral aid from the so-called like-minded donors, Denmark, the Netherlands, Norway, and Sweden. Often are held up as model donors whose aid allocations more closely mirror need and aid effectiveness, these small countries have very limited ability to pressure the World Bank to serve narrow geopolitical objectives (though they may have more influence in defining best practices). Thus, like-minded donor aid should serve as a proxy for elements of need and aid effectiveness not captured by other variables already included. If we replace U.S. bilateral aid with bilateral aid from the like-minded donors, the estimated hazard ratio is again greater than one with a similar size and significance level. If we include both bilateral aid variables simultaneously, we cannot reject that the effects are the same (though neither effect is estimated precisely). Thus, the most straightforward interpretation is that our other control variables do not sufficiently capture need/effectiveness so that bilateral aid (whether from the like-minded donors, the U.S., or the G7) serves as a proxy for need/effectiveness rather than measuring geopolitical influence in a meaningful fashion. An alternative interpretation—similarly unrelated to geopolitical issues—is that IEG's PPARs might be most useful if they are carried out in countries that also receive aid from many bilateral donors. As major bilateral donors to a country, World Bank shareholders might be more interested in the PPARs' outcomes and these outcomes could be used for learning beyond the World Bank.

All the other geopolitical variables—U.S. military aid, UN voting, UNSC nonpermanent membership, and World Bank Executive Board membership—appear to have no impact on the PPAR hazard rate. None of the hazard ratios are statistically different from one and most are quantitatively very close to one. In short, Table 4 provides little evidence that political economy factors play a role in PPAR risk though a slightly more nuanced picture emerges below.

Do rating revisions delay PPARs?

One factor which could delay the release of a PPAR is a change in the rating the project receives. Delays could happen if such cases are more complex, if changing a rating requires more deliberation within IEG, if feedback from Operations or the borrowing government takes longer or as a result of external pressure.[23] This issue can only be examined for cases with PPARs, of course. To explore this question, we define *Downgrade* as a project with a "Satisfactory" binary ICR rating and a subsequent "Unsatisfactory" binary PPAR rating and *Upgrade* as a project with an "Unsatisfactory" binary ICR rating and a subsequent "Satisfactory" binary PPAR rating.

Note that "Satisfactory" ICR ratings outnumber "Unsatisfactory" ICR ratings in the overall estimation sample (73% v. 27%); for the PPAR sample, this pattern is even more pronounced (79% v. 21%) since IEG selects a disproportionate number of projects with "Satisfactory" ICR ratings.[24] The result is far more downgrades (157) than upgrades (45) in our sample. However if we compare the incidence of downgrades and upgrades, the imbalance disappears: P(*Downgrade* | ICR = "Satisfactory") = 14.6% and P(*Upgrade* | ICR = "Unsatisfactory") = 15.4%.

[Figure 6 about here]

Returning to the question of whether rating changes are associated with longer delays between ICR and PPAR (the *PPAR lag*), Figure 6 provides separate *PPAR lag* distributions by rating change (no change, *Downgrade*, *Upgrade*). As the figure illustrates, the time between ICR

[23] IEG describes the process as follows: "Each PPAR is subject to internal IEG peer review, Panel review, and management approval. Once cleared internally, the PPAR is commented on by the responsible Bank department. The PPAR is also sent to the borrower for review. IEG incorporates both Bank and borrower comments as appropriate, and the borrowers' comments are attached to the document that is sent to the Bank's Board of Executive Directors. After an assessment report has been sent to the Board, it is disclosed to the public." (IEG 2015A, vi)

[24] For comparison, the ratio is 71% v. 29% for projects with no PPAR.

and PPAR is substantially longer when there is a rating change than when there is not; the mean difference is about one year. The pattern looks similar for both downgrades and upgrades.[25]

Table 5 presents hazard model estimation results where we restrict the sample to the 1371 cases with PPARs (uncensored cases) to examine the role of rating changes. Column 1 repeats the specification from Column 3 of Table 3 for comparison. Column 2 adds the *Downgrade* and *Upgrade* variables to the baseline specification; these are implicitly compared against the case of no rating change. Both enter with hazard ratios less than one and of similar magnitude but only the hazard ratio for *Downgrade* is statistically significantly different from one, perhaps due to the relative scarcity of upgrades. Other hazard ratio estimates are essentially unaffected.

Column 3 introduces *WBEB*, the dummy variable indicating membership on the World Bank's Executive Board. Of the geopolitical variables considered thus far, only *WBEB* enters the specification with a statistically significant hazard ratio in this sample (and even then, only at the 10% level). Column 3, however, reports a slightly more sophisticated specification that also includes the interaction of the Board membership dummy with the *Downgrade* and *Upgrade* dummies. In this specification, the hazard ratio for *WBEB* × *Downgrade* is significantly less than one while the hazard ratio for *WBEB* × *Upgrade* is significantly greater than one. For the uninteracted terms, the hazard ratio for *Downgrade* continues to be significantly less than one and the hazard ratio for *Upgrade* is now marginally significantly less than one while *WBEB* is not significant. Taken together, these results imply: 1) PPARs that downgrade ratings experience delays which are longer when the borrowing country occupies an important position in the World

[25] More sophisticated breakdowns (e.g., taking into account whether the project is part of a multi-project PPAR and, if so, where in the sequence of projects it is) produce similar results. That is, the time between ICR and PPAR is substantially longer when there is a rating change than when there is not.

Bank governance structure; 2) PPARs that upgrade ratings are also delayed except when the borrowing country occupies an important position in the World Bank in which case the PPAR is instead accelerated; and 3) when PPAR findings are unremarkable (i.e., no rating change), the borrowing country's importance within the World Bank does not impact the release of these findings. This pattern suggests both a bureaucratic procedure (e.g., another layer of review) when PPARs change ratings but also a response to power within the institution, delaying bad news and expediting good news when countries hold important positions in the World Bank's governance structure.

The final column of Table 5 introduces dummy variables for non-permanent UNSC membership, at the time of the ICR and at the time of the PPAR. Note that, like *Downgrade* and *Upgrade*, *UNSC* (*PPAR*) can only be examined for cases where there is a PPAR. As in the full sample presented in Table 4, the hazard ratio for *UNSC* (*ICR*) is extremely close to one and far from statistical significance. In contrast, the hazard ratio for *UNSC* (*PPAR*) is significantly less than one. In the case that a country is a non-permanent member of the UNSC at the time of the PPAR, the PPAR is significantly delayed.[26]

4. PPAR Outcome Ratings

[26] We do not interact *UNSC* with *Downgrade* or *Upgrade* as the number of *UNSC* members is too small and so results would be very sensitive to sampling variation (outliers). For example, there are only 7 cases where a UNSC member has a downgrade. Looking at UNSC membership, timing is critically important (since nonpermanent membership lasts only 2 years) where as it is less critical when considering other geopolitical variables (bilateral aid, UN voting alignment, etc.) which vary less over time. This explains why results differ between *UNSC* (*ICR*) and *UNSC* (*PPAR*) while results for the other geopolitical variables do not depend on whether measured at the time of the ICR or the PPAR.

Understanding the selection of projects for PPARs explored in the hazard analysis above is important for understanding differences between ICR and PPAR ratings. We now turn to examining what factors influence IEG's evaluation process and how do the resulting ratings differ from the ICR ratings. In particular, we ask if bureaucratic factors or geopolitics influence rating by the Independent Evaluation Group or is IEG sufficiently insulated from such outside pressures.

A few simple graphs for variables that proved significant in the earlier hazard model provide an interesting starting point for this analysis. Figure 7 explores the bivariate relationship between rating changes and non-permanent UNSC membership. For each project with a PPAR, we subtract the binary ICR rating from the binary PPAR rating, resulting in a −1 (downgrade), 0 (no change) or +1 (upgrade). Then for each year with available data, we calculate the average of this difference for projects in countries that are non-permanent UNSC members that year and for other projects. We plot each resulting point with UNSC values on the Y-axis and non-UNSC values on the X-axis. Figure 7 includes each point as a circle proportional to the number of underlying projects to give a sense of where the mass of the data lies (since the number of PPARs varies by year). If there were no relationship between non-permanent UNSC membership and rating changes, the data points should cluster symmetrically about the 45-degree line.

[Figure 7 about here]

Figure 7 has two interesting features. First, the data point for FY2002 (July 1, 2001-June 30, 2002) is a clear outlier. Given the terrorist attacks on September 11, 2001, this is perhaps unsurprising, though why this episode would be linked to rating downgrades for UNSC members' projects is unclear. Second, rather than clustering symmetrically about the 45-degree line, the bulk of the data lie above the line. This indicates that the ratings for projects of UNSC non-permanent

members are less likely to be downgraded/more likely to be upgraded by IEG than are the ratings for projects of other borrowers. This pattern is even more pronounced if we exclude FY2002.

[Figure 8 about here]

Figure 8 repeats this exercise but for World Bank Executive Board membership. In contrast to the previous graph, the data points fall relatively symmetrically about the 45-degree line. While three data points stand slightly apart from the rest (FY1990, FY1997, and FY2000), there are no obvious outliers as in the previous graph. In short, these two graphs suggest that external geopolitical importance might influence IEG ratings while internal bureaucratic power might not.

We explore the determinants of PPAR ratings more systematically by estimating a PPAR rating equation of the following form:

$$PPAR_{ij} = \beta_0 + \beta_1 ICR_{ij} + \boldsymbol{\beta}_2' \boldsymbol{X}_{ij} + \boldsymbol{\beta}_3' \boldsymbol{Z}_{ij} + \varepsilon_{ij} \quad (1)$$

where ICR_{ij} is the ICR rating for project i in country j, $\boldsymbol{X}_{ij}$ is a matrix that includes project characteristics as well as country characteristics measured at the relevant time for project i, and $\boldsymbol{Z}_{ij}$ is a matrix that includes bureaucratic or geopolitical factors describing country j at the relevant time for project i. We estimate this three ways. Using binary PPAR and ICR ratings, we estimate equation (1) via probit and via OLS. Using the more fine-grained 6-point PPAR and ICR ratings available from 1994 on, we again estimate equation (1) via OLS but replace ICR_{ij} with dummy variables ($ICR2$ to $ICR6$) for the individual rating values where $ICR1$ ("Highly Unsatisfactory") is the excluded category.[27]

[27] We do not present ordered probit results for the six-point rating scale because of the challenges involved in presenting useful marginal effects in this setting. However, the sign and significance of estimated coefficients for the latent variable model in the ordered probit match those in the linear model. Note that all specifications have PPAR ratings as the dependent variable and ICR ratings as one of the explanatory variables. This is a more flexible specification than using the

[Table 6 about here]

Table 6 presents estimation results and marginal effects for non-permanent UNSC membership. The first four columns report results for a parsimonious specification that includes only the ICR rating and non-permanent UNSC membership at project approval, at ICR and at PPAR. Column 1 reports probit estimation results while Column 2 presents marginal effects. Column 3 reports OLS results using the binary version of the PPAR and ICR ratings. Column 4 repeats that but using the 6-point PPAR rating as the dependent variable and the set of ICR dummies as independent variables. The last four columns mirror the first four but for our "full specification" that includes country fixed effects, approval year dummies, PPAR year dummies, project characteristics, and a set of time-varying country characteristics. We include these more complete specifications to check for omitted variable bias but the vast majority of the additional variables prove insignificant and we do not report their coefficient estimates in the table. Based on the pattern identified in Figure 7, we exclude PPARs completed in FY2002.[28]

A very consistent pattern emerges from Table 6. As one would expect, ICR ratings are strongly predictive of PPAR ratings; a satisfactory ICR rating increases the chances of a satisfactory PPAR rating by 68 to 77 percentage points, ceteris paribus. If a country holds a non-

change in rating as the dependent variable and avoids complications in the probit setting. Nonetheless, results in the OLS estimations are largely the same if we do use the change in rating as the dependent variable.

[28] Project characteristics include an IDA dummy and a Program Loan dummy. Country characteristics include GDP per capita, GDP growth rate, inflation, openness, the ratio of short-term debt to total debt, the average number of years of schooling, the combined Freedom House rating, and government years in office, all at approval. The sample size falls from 1500 observations (on 120 countries) in the parsimonious specifications to 1329 observations (on 93 countries) in the full specification due to missing data. This is less than the 1371 observations in Table 5 because we exclude FY2002 PPARs. The full specification probit sample is slightly smaller still, at 1219 observations (on 69 countries) due to lack of variation in PPAR ratings (all Satisfactory or all Unsatisfactory) in some countries.

permanent seat on the UNSC at the time of the PPAR rating, the chances of a satisfactory rating increase by 6.4 to 9.8 percentage points. In contrast, earlier membership (at approval or ICR) does not directly influence PPAR ratings. Dreher *et al*. (2013) and Kilby (2015) find evidence of a negative effect of non-permanent UNSC membership at the time of approval in certain settings; however, the included ICR rating should capture such effects. We might expect a negative coefficient on *UNSC* (*ICR*) if PPAR ratings correct for a bias in the ICR ratings; the point estimates are negative in 5 of 6 specifications but never approach significance.

[Table 7 about here]

Table 7 presents some evidence about what drives these findings. We define World Bank debt dummies for the bottom 25% (*Low Debt*), middle 50% (*Medium Debt*) and top 25% (*High Debt*) and then interact these variables with *UNSC* (*PAR*) to determine if a subset of cases drives the results in Table 6. This does indeed appear to be the case: the estimated coefficient on *UNSC* (*PAR*) is consistently significant—and generally larger—only for the *High Debt* case. The probit, for example, indicates that the probability of a satisfactory PPAR rating is 13 to 15 percentage points higher for countries that owe a lot to the World Bank when they serve as non-permanent members of the UNSC at the time of the PPAR rating as compared to cases where they do not serve on the UNSC. The point estimates for the interaction term with medium World Bank debt are also positive but significant or marginally significant in only half the cases. The point estimates for the interaction term with low World Bank debt are negative but never significant. We also explored other measures that the literature suggests might condition the effects of geopolitics in international financial institutions (Stone 2008, Dreher *et al*. 2013) including the ratio of short term debt to total debt and the debt service ratio. None of the categories (low, medium, high) associated with these variables identify the cases driving the UNSC results.

World Bank debt, however, is closing linked to country size. The correlation between population and World Bank debt in our estimation sample is 0.9. Estimation results partitioning project observations into low, medium, and high population levels (i.e., reproducing Table 7 but based on population rather than World Bank debt) are essentially the same. Thus, we cannot separately identify whether UNSC membership matters for countries that are large borrowers or large countries that are borrowers.

In a series of estimations not reported here, we repeated the exercise in Table 6 with the other geopolitical variables introduced for the PPAR hazard analysis in Table 4 (U.S. and G7 bilateral aid, U.S. military aid, U.S. and G7 UN voting alignment, World Bank Executive Board membership). We did not find robust effects on PPAR ratings for any of these, whether based on their values at the time of project approval, ICR or PPAR. Furthermore, interactions with low, medium, and high levels of World Bank Debt or the short-term debt to total debt ratio (parallel to Table 7) did not identify any categories with robust effects.

Finally, we explored whether selection issues impact our results. We estimate a Heckman selection model where the selection equation is a probit using the same specification as Column 2 from Table 3 and the PPAR outcome equation is either the linear probability model (Columns 3 or 7 from Table 6) or the 6-point rating (Columns 4 or 8 from Table 6). This set-up lends itself well to a selection model since there are a number of variables which influence selection but have no theoretical link to outcomes (ICR quality, number of World Bank projects finishing in the same year, and June ICR). Whether or not we can reject the independence of the error terms in the two equations (i.e., whether or not the selection model is relevant) depends on which variables we include. However, in all cases the estimated coefficient for *UNSC* (*PPAR*) remains positive and

statistically significant with little variation in magnitude. Thus, selection effects, if any, do not materially influence our findings.

5. Conclusion

This paper investigates two functions of the World Bank's Independent Evaluation Group. We first examine how IEG selects projects for a Project Performance Assessment Report, a comprehensive review carried out for about a quarter of projects that have completed implementation. We then turn to how PPAR ratings done by IEG staff differ from the initial ratings that Operations staff submit as part of their Implementation Completion Report.

IEG makes clear that the PPAR selection process is not random. IEG focuses on projects with positive initial ratings; in fact, the higher the rating the more likely the project is to be reviewed. In addition to this strategy of "learning from success," IEG also samples more heavily from the 6% of projects it identifies as having poorly done ICRs. IEG is more like to select large projects and projects from less developed countries for additional review in a PPAR. Program loans, especially those in high inflation environments, are also more likely to subject to a PPAR than investment projects. There is a surge of Implementation Completion Reports filed at the end of the World Bank's fiscal year; these June ICRs are individually less likely to be selected for a PPAR as IEG attempts to allocate its scarce resources. When multiple projects in a single country are completed in the same year, these projects are more likely to be selected for a PPAR as IEG staff undertake "cluster audits" to economize on travel expenses and time. All these patterns seem well in-line with IEG's mission.

There are, however, some elements of PPAR selection that are less obviously in-line with the mission. IEG is more likely to review projects in countries with easier working conditions

such as well-established, democratic governments. IEG is also more likely to conduct a PPAR—which includes a country visit by IEG staff and consultants—and conduct it sooner rather than later if the country is a popular tourist destination. Institutional power appears to impact the PPAR release date. PPARs that result in rating changes on average take an additional year. If the borrowing country is on the World Bank Executive Board and the rating change is a downgrade, the release of this bad news is delayed twice as along. Conversely, if the borrowing country is on the World Bank Executive Board and the rating change is an upgrade, IEG releases the PPAR more quickly than if there had been no rating change. PPAR releases also take longer if the borrowing country is a non-permanent member of the UNSC although this delay does not appear related to whether the rating changes.

We also find some evidence linking IEG ratings to geopolitics. Controlling for the previous ICR rating, the PPAR rating is more likely to be positive if the borrowing country holds one of the non-permanent seats on the UN Security Council. This result is driven primarily by higher ratings for countries that both occupy a non-permanent UNSC seat and are important World Bank customers (i.e., are large World Bank borrowers). Such countries are also large in terms of population and, because of a high correlation between population size and borrowing, our approach cannot separately identify the two effects. While these results are robust, we find no robust links to a range of other variables used in the aid literature to capture geopolitics.

This paper does not explain why these various patterns exist. Some of the patterns in PPAR selection may simply reflect a natural bureaucratic tendency toward a more pleasant work environment and may have little consequence in terms of the effectiveness of the organization. Delaying the release of "bad news" (rating downgrades) for institutionally important Board members is perhaps more troubling but there is no evidence Board member evaluation privileges

extend beyond this. That is, it appears that World Bank Executive Board members can delay or expedite a PPAR (or that World Bank staff do this on behalf of EDs or for their own reasons) but we do not find convincing evidence that Executive Board members prevent PPARs or alter their outcome ratings. Furthermore, we do not have evidence that explains why borrowing governments would want to delay or expedite a PPAR release and can only speculate that PPAR ratings might influence elections or access to additional aid resources.

In contrast, geopolitical importance does have a measurable impact on project outcome ratings done by IEG. In this case, the likely channel of influence is the U.S. or G7 indirectly applying pressure to promote favorable project ratings. Again, the tangible benefits from higher project ratings might be political or financial but exactly what they are remains an open question for future research.

References

Azevedo, J.P. 2011. wbopendata: Stata module to access World Bank databases, Statistical Software Components S457234, Boston College Department of Economics. http://ideas.repec.org/c/boc/bocode/s457234.html

Beck, Thorsten, George Clarke, Alberto Groff, Philip Keefer, and Patrick Walsh. 2001. New tools in comparative political economy: The Database of Political Institutions. World Bank Economic Review 15:165-176.

Blum, Jürgen René. 2014. What factors predict how public sector projects perform? A review of the World Bank's public sector management portfolio. World Bank Policy Research Working Paper 6798.

Bulman, David, Walter Kolkma, and Aart Kraay. 2015. Good countries or good projects? Comparing macro and micro correlates of World Bank and Asian Development Bank project performance. World Bank Policy Research Working Paper 7245.

Buntaine, Mark, and Bradley C. Parks. 2013. When do environmentally-focused assistance projects achieve their objectives? Evidence from World Bank post-project evaluations. Global Environmental Politics 13(2):67-90.

Chauvet, Lisa, Paul Collier, and Marguerite Duponchel. 2010. What explains aid project success in post-conflict situations? World Bank Policy Research Working Paper 5418.

Chauvet, Lisa, Paul Collier, and Andreas Fuster. 2015. Supervision and project performance: A principal-agent approach. DIAL Working Paper 2015-04.

Cruz, Cesi, and Philip Keefer. 2013. The organization of programmatic parties and the politics of bureaucratic reform. World Bank Policy Research Working Paper 6686.

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.

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 Victoria Levin. 2005. Sowing and reaping: Institutional quality and project outcomes in developing countries. World Bank Policy Research Working Paper # 3524.

Dollar, David and Jakob Svensson. 2000. What explains the success or failure of structural adjustment programmes? The Economic Journal 110(466):894-917.

Dreher, Axel, Jan-Egbert Sturm and James Vreeland. 2009. Development aid and international politics: does membership on the UN Security Council influence World Bank decisions? Journal of Development Economics 88: 1-18.

Dreher, Axel, Stephan Klasen, James Raymond Vreeland, Eric Werker. 2013. The costs of favoritism: Is politically-driven aid less effective? Economic Development and Cultural Change 62(1):157-191.

Fleck, Robert K., and Christopher Kilby. 2010. Changing aid regimes? U.S. foreign aid from the Cold War to the War on Terror. Journal of Development Economics 91(2):185-197.

Freedom House. 2012. Freedom in the World.

Geli, Patricia, Aart Kraay, and Hoveida Nobakht. 2014. Predicting World Bank project outcome ratings. World Bank Policy Research Working Paper 7001.

Girod, Desha M., and Jennifer Tobin. 2011. Take the money and run: The determinants of compliance with aid agreements. Presented at the International Political Economy Society, Madison, Wisconsin, November 2011.

Grasso, Patrick G., Sulaiman S. Wasty, Rachel V. Weaving (Eds). 2003. World Bank Operations Evaluation Department: The First 30 Years. Washington, DC: The World Bank.

Guillaumont, Patrick, and Rachid Laajaj. 2006. When instability increases the effectiveness of aid projects. World Bank Policy Research Working Paper # 4034.

IEG (World Bank Independent Evaluation Group). 2011. IEG Work Program and Budget (FY12) and Indicative Plan (FY13-14). Washington, DC: The World Bank.

IEG (World Bank Independent Evaluation Group). 2014. Independent Evaluation Group - World Bank Project Performance Ratings. http://ieg.worldbankgroup.org/ratings

IEG (World Bank Independent Evaluation Group). 2015A. Project Performance Assessment Report. Brazil: First Programmatical Reform Loan for Environmental Sustainability. Report No.: 92270-BR.

IEG (World Bank Independent Evaluation Group). 2015B. World Bank Project Performance Ratings - Codebook. https://ieg.worldbankgroup.org/Data/reports/ieg-wb-project-performance-ratings-codebook_092015.pdf

Isham, Jonathan, Daniel Kaufmann, and Lant Pritchett. 1997. Civil liberties, democracy, and the performance of government projects. World Bank Economic Review 11(2):219-242.

Isham, Jonathan, and Daniel Kaufmann. 1999. The forgotten rationale for policy reform: The productivity of investment projects. Quarterly Journal of Economics 114(1):149-184.

Kaufmann, Daniel, and Yan Wang. 1995. Macroeconomic policies and project performance in the social sectors: A model of human capital production and evidence from LDCs. World Development 23(5)751-765.

Kaja, Ashwin and Eric Werker. 2010. Corporate governance at the World Bank and the dilemma of global governance. World Bank Economic Review 24(2):171-198.

Kilby, Christopher. 2009. The political economy of conditionality: an empirical analysis of World Bank loan disbursements. Journal of Development Economics 89(1):51-61.

Kilby, Christopher. 2013A. An empirical assessment of informal influence in the World Bank. Economic Development and Cultural Change 61(2):431-464.

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

Kilby, Christopher. 2015. Assessing the impact of World Bank preparation on project outcomes. Journal of Development Economics 115:111-123.

Limodio, Nicola. 2011. The success of infrastructure projects in low income countries and the role of selectivity. World Bank Policy Research Working Paper # 5694.

Malesa, Thaddeus, and Peter Silarszky. 2005. Does World Bank effort matter for success of adjustment operations? In Stefan Koeberle, Harold Bedoya, Peter Silarsky, and Gero Verheyen (Eds.), Conditionality Revisited: Concepts, Experiences, and Lessons. (Washington: The World Bank). Chapter 13, pages 127-141.

Malik, Rabia, and Randall W. Stone. 2016. Corporate influence in World Bank lending. University of Rochester, Department of Political Science Working Paper.

Michaelowa, Katharina, and Axel Borrmann. 2006. Evaluation bias and incentive structures in bi- and multilateral aid agencies. Review of Development Economics 10(2):313-329.

Moll, Peter, Patricia Geli, and Pablo Saavedra. 2015. Correlates of success in World Bank development policy lending. World Bank Policy Research Working Paper 7181.

Morrison, Kevin M. 2013. Membership no longer has its privileges: The declining informal influence of board members on IDA lending. Review of International Organizations 8(2):291-312.

OECD (Organization for Economic Cooperation and Development). 2015. Geographical distribution of financial flows: Flows to developing countries. OECD International Development Statistics (database). DOI: http://dx.doi.org/10.1787/data-00566-en

Pohl, Gerhard, and Dubravko Mihaljek. 1992. Project evaluation and uncertainty in practice: A statistical analysis of rate-of-return divergences of 1,015 World Bank projects. The World Bank Economic Review 6(2):255-277.

Ralston, Laura. 2014. Success in difficult environments: A portfolio analysis of fragile and conflict-affected states. World Bank Policy Research Working Paper 7098.

Smets, Lodewijk, Stephen Knack, and Nadia Molenaers. 2013. Political ideology, quality at entry and the success of economic reform programs. Review of International Organizations 8(4):447-476.

Stone, Randall. 2008. The scope of IMF conditionality. International Organization 62:589-620

Strezhnev, Anton and Erik Voeten. 2012. United Nations General Assembly Voting Data, http://hdl.handle.net/1902.1/12379 UNF:5:fWzDBiI+iY41v5zJF3JnoQ== Erik Voeten [Distributor] V4 [Version]

Sud, Inder, and Jane Olmstead-Rumsey. 2012. Development outcomes of World Bank projects: Real or illusory improvements. George Washington University Institute for Global and International Studies Working Paper #1.

United Nations. 2014. Countries Elected Members of the Security Council. http://www.un.org/en/sc/members/elected.asp Accessed 9/15/2014.

USAID (United States Agency for International Development). 2014. U.S. overseas loans and grants: obligations and loan authorizations, July 1, 1945-September 30, 2012 (Greenbook). https://explorer.usaid.gov/reports-greenbook.html Accessed 3/20/2014.

U.S. State Department. 1984-2013. Voting Practices in the United Nations. GPO, Washington.

Winters, Matthew S. 2014. Targeting, accountability and capture in development projects. International Studies Quarterly 58(2):393-404.

World Bank. 2012A. Access to Information. Annual Report FY 2011. Moving Forward Transparency and Accountability. (Washington: The World Bank).

World Bank. 2012B. Projects database. http://go.worldbank.org/KPMUDAVVT0 Accessed 11/30/2012.

World Bank. 2014. World Development Indicators. http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators Accessed 10/14/2014.

Figure 1: Project Timeline. Adapted from Geli *et al*. (2014). Not to scale.

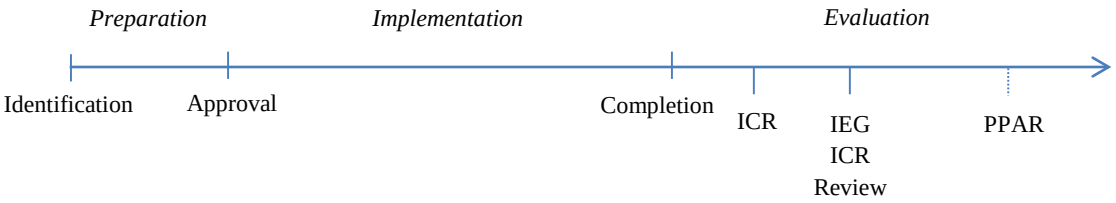


Figure 2: First Evaluation Type by Evaluation Fiscal Year

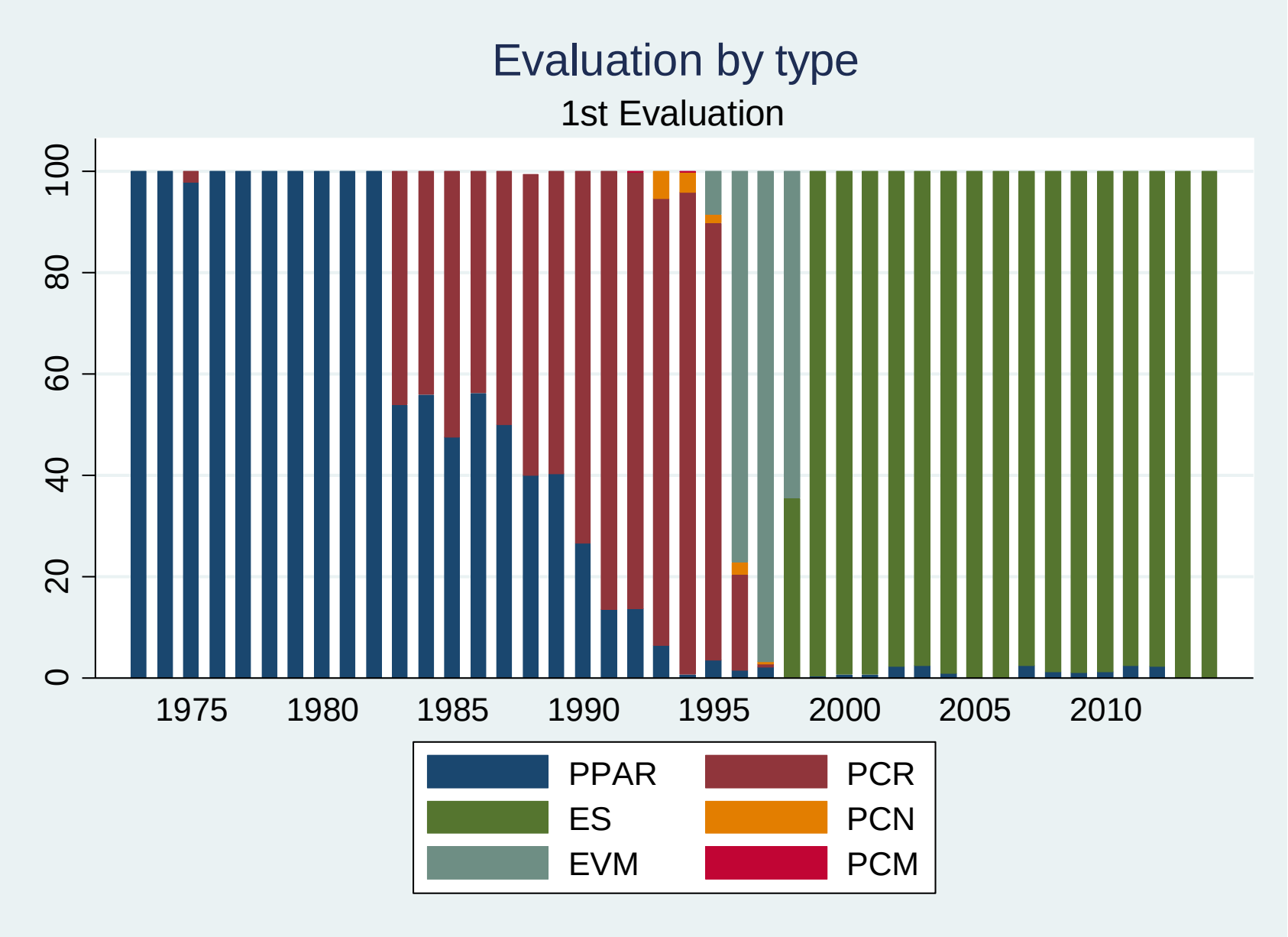


Figure 3: Cumulative Hazard Rate Functions for Selected Variables.

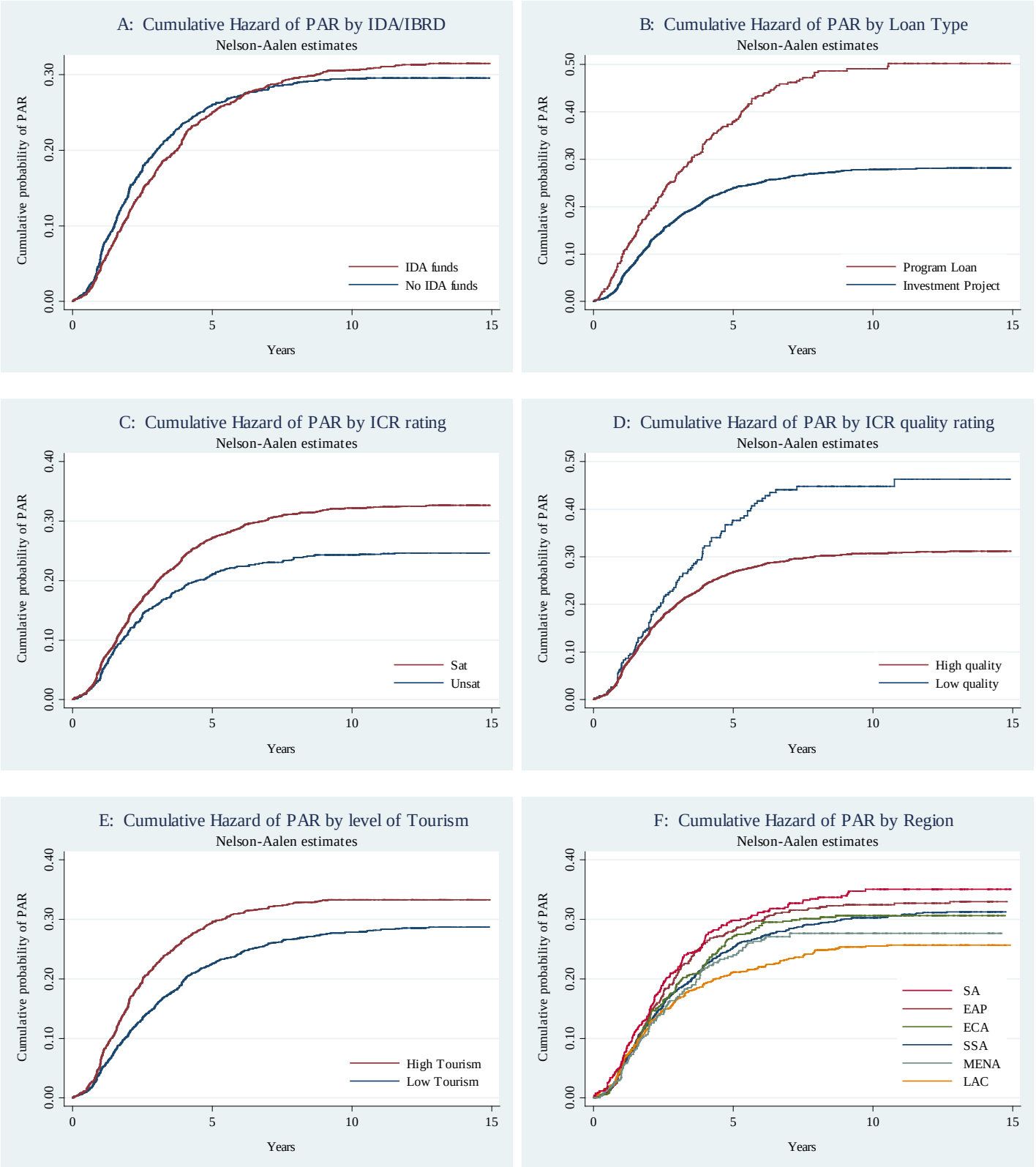


Figure 4: Time between ICR and PPAR

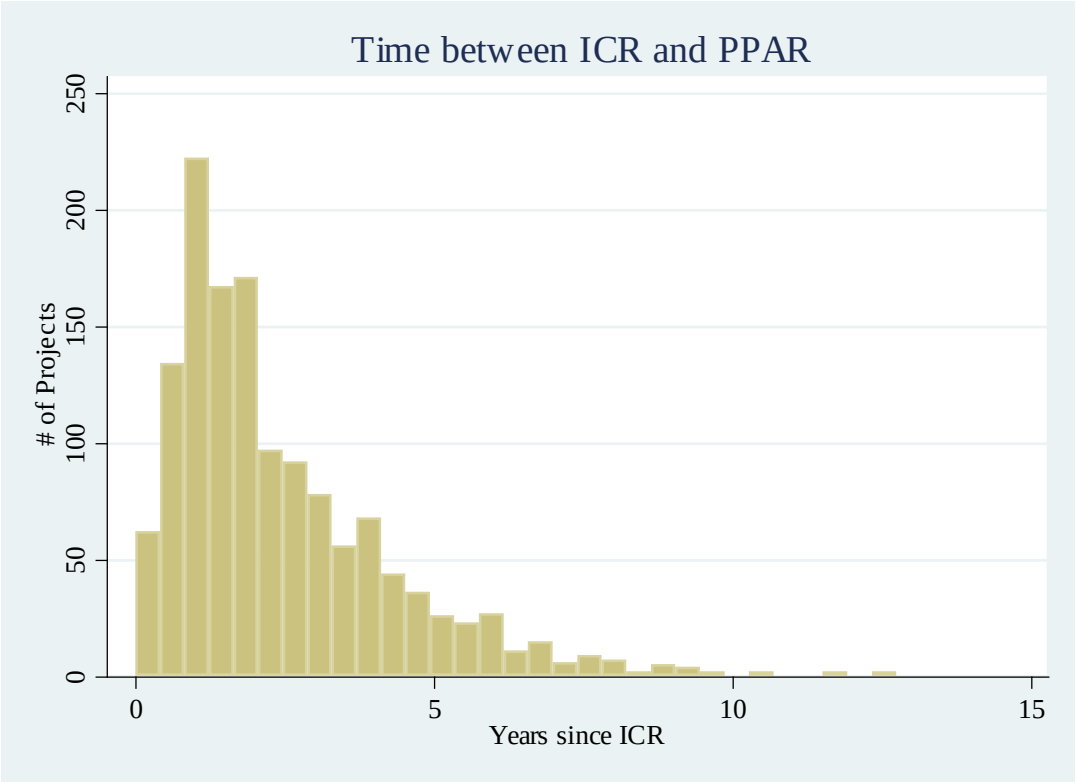


Figure 5: Time since ICR for projects without a PPAR
(as of September 30, 2013)

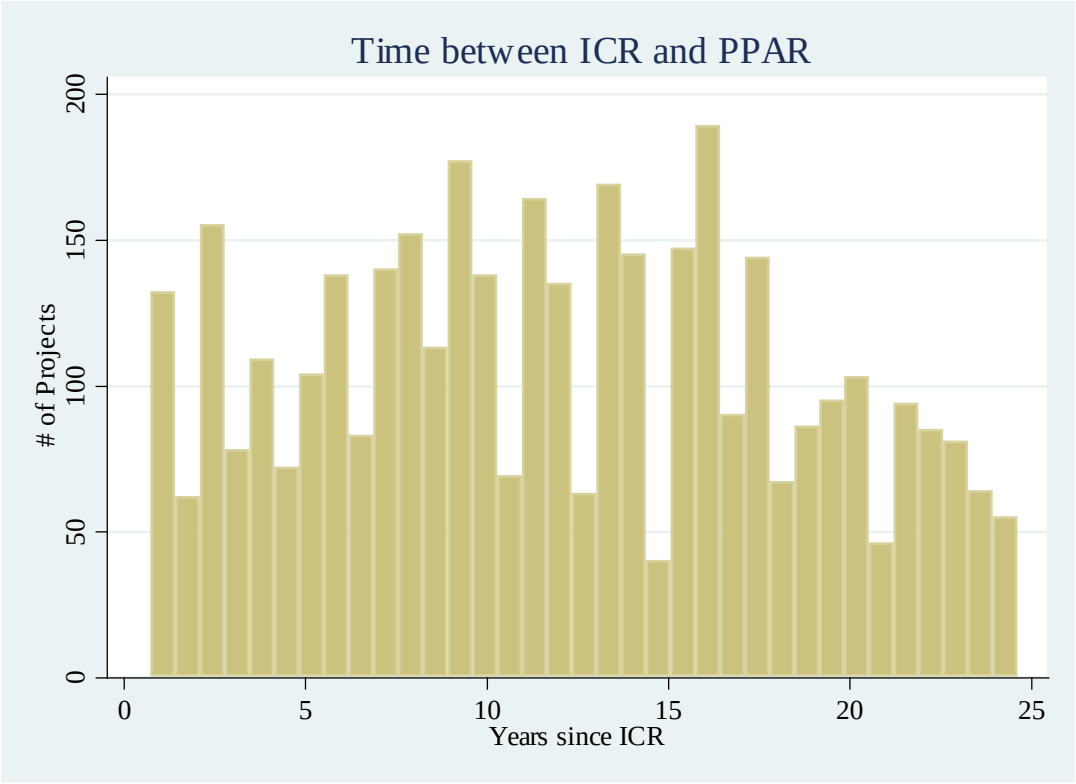


Figure 6: *PPAR lag* by outcome rating change

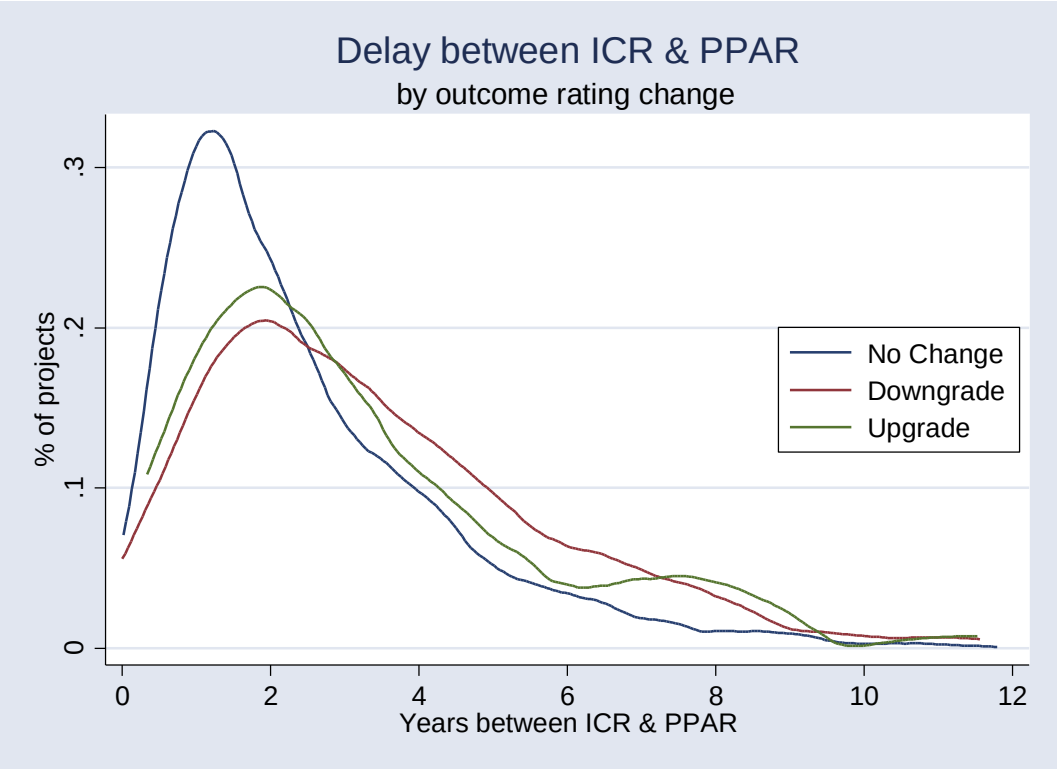
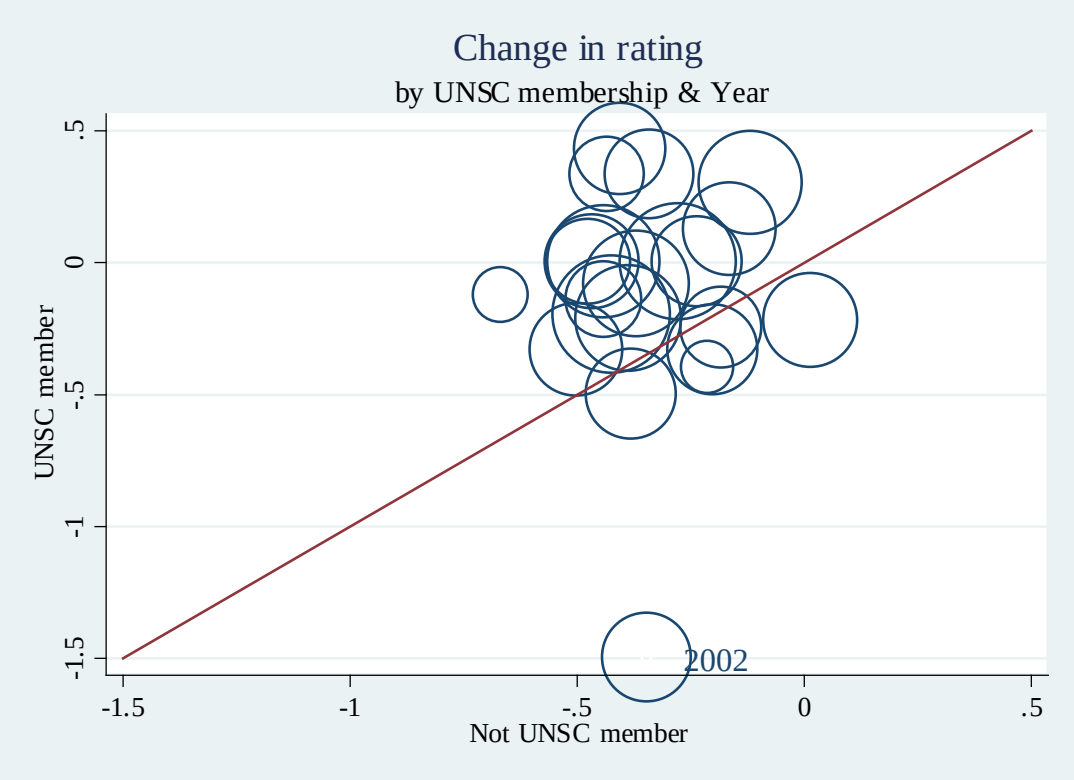


Figure 7: Rating change and non-permanent UNSC membership



F

Figure 8: Rating Change and World Bank Executive Board Membership

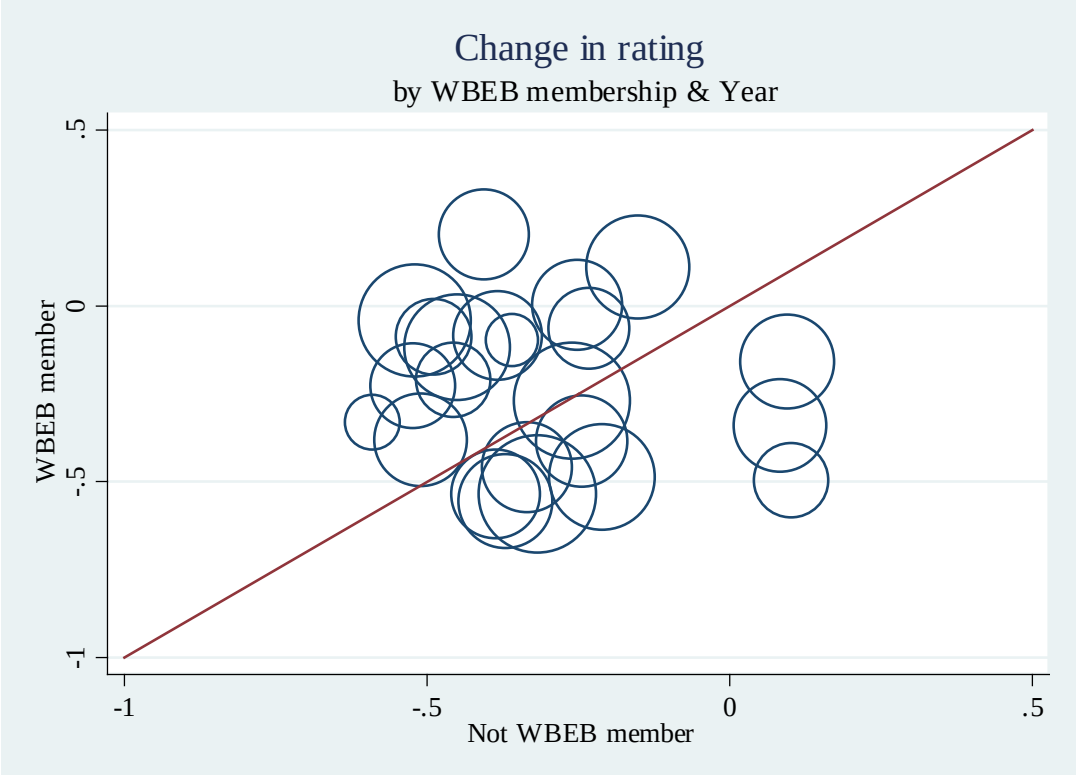


Table 1: IEG Evaluation Ratings

Rating	Coverage[A]	Scale[B]	Description
Overall Performance			
Outcome	1972-1992 1993 1994-	Binary 4 point 6 point[C]	"extent to which the operation's major relevant objectives were achieved, or are expected to be achieved, efficiently" (IEG 2015B, 14)
Sustainability	1989-2000 2000.5-2008	3 point 5 point	"At the time of the evaluation, the resilience to risk of future net benefits flows." (IEG 2015B, 29)
Risk To Development Outcome (gradually replaced Sustainability)	2006.5-	4 point	"risk, at the time of evaluation, that development outcomes (or expected outcomes) will not be maintained (or realized)" (IEG 2015B, 16)
Institutional Development Impact	1989-1999 1999.5-2007	3 point 4 point	"extent to which a project improves the ability of a country or region to make more efficient, equitable and sustainable use of its human, financial, and natural resources" (IEG 2015B, 27)
World Bank Performance			
Quality at Entry	1994-2006 2006.5-	4 point 6 point	"extent to which the Bank identified, facilitated preparation of, and appraised the operation such that it was most likely to achieve planned development outcomes and was consistent with the Bank's fiduciary role" (IEG 2015B, 17)
Quality of Supervision	1991-1992 1993-2006 2006.5-	Binary 4 point 6 point	"extent to which the Bank proactively identified and resolved threats to the achievement of relevant development outcomes and the Bank's fiduciary role" (IEG 2015B, 18)
Overall Bank Performance	1997-2006 2006.5-	4 point 6 point	"The quality at entry and quality of supervision ratings should be combined into a rating of overall Bank Performance." (IEG 2015B, 19)
Borrower Performance			
Borrower Preparation	1993-2003	4 point	"government/implementing agency performance on the preparation of this project. Consider specifically whether the government/implementing agency took account of economic, financial, technical, policy, and resource considerations, and ensured participation of major stakeholders in preparing the project" (IEG 2015B, 30)
Government Compliance	1993-2006	4 point	Renamed Government Performance
Government Performance	2006.5-	6 point	"extent to which the borrower ... government ... ensured quality of preparation and implementation, and complied with covenants and agreements, towards the achievement of development outcomes" (IEG 2015B, 21)
Implementing Agency Performance	1993-2006 2006.5-	4 point 6 point	"extent to which the borrower...implementing agency or agencies...ensured quality of preparation and implementation, and complied with covenants and agreements, towards the achievement of development outcomes" (IEG 2015B, 21)
Overall Borrower Performance	1997-2006 2006.5-	4 point 6 point	"The ratings for government performance and implementing agency or agencies' performance should be combined into a rating of Borrower Performance, per guidance below." (IEG 2015B, 22)
Evaluation Process			
Quality of ICR	1983-1989 1990-1996 1997-	Binary 4 point 3 point	IEG evaluator assessment of PCR/ICR quality. Prior to 1997, very few projects have ratings (2% prior to 1990, 14% 1990-1996). After 1996, 80% of projects rated. See IEG (2015B, 24).
Quality of M&E	2006.5-	4 point	"based on an assessment of three main elements: (i) M&E design; (ii) M&E implementation; and (iii) use of M&E data... The evaluator is asked to discuss separately each of the three elements of M&E quality and to arrive at an overall quality of M&E rating on a 4-point scale." (IEG 2015B, 25)

[A] Reflects effective dates in rating data rather than official dates of policy changes; See IEG (2015B) for official dates. A suffix of .5 indicates change on or after July 1 (start of fiscal year); otherwise, calendar years.[B] Ratings also include "NOT AVAILABLE," "NOT APPLICABLE," "NOT RATED," and "NON-EVALUABLE" as well as missing values.
[C] Database includes two "MARGINALLY SATISFACTORY" and three "MARGINALLY UNSATISFACTORY" ratings. These are not officially recognized designations (IEG 2015B).
See Sud and Olmstead-Rumsey (2012) and IEG (2015B) for detailed discussions of the evolution of ratings and its implications for aggregate reporting.

Table 2: Descriptive Statistics for Hazard Rate Estimation

Variable	N	Mean	SD	Min	Max	Description
Outcome (ICR)	5155	0.73	0.44	0	1	=1 if project rated "Satisfactory"
IDA	5155	0.54	0.50	0	1	=1 if project included IDA funding
ICR quality	5155	0.94	0.24	0	1	=1 if ICR quality rated "Satisfactory" [A]
log World Bank debt	5155	0.68	1.51	-4.75	3.53	log of debt to IBRD/IDA in billions of 2005 USD [B,C]
log Project size	5155	4.00	1.21	-0.63	8.00	log of loan amount in 2005 USD (millions)
log # WB projects	5155	1.40	0.77	0	3.04	log of # ICRs completed same fiscal year (by country)
June ICR	5155	0.18	0.38	0	1	=1 if ICR published in June (end of FY)
Tourism	5155	13.98	1.81	9.53	17.51	log of Tourist Arrivals (country average)
Years in office	5155	7.30	7.67	1	45	Executive years in office [B]
Freedom House	5155	4.07	1.50	1	7	Average of political rights & civil liberties (1 is best) [B,C]
log Population	5155	17.10	1.87	11.48	21.02	log of population
log GDP PC	5155	7.00	1.04	4.78	9.31	log of GDP per capita in 2000 USD [B,C]
GDP growth	5155	4.61	3.43	-17.49	28.65	GDP growth rate [B,C]
Inflation	5155	0.16	0.44	-0.13	8.07	Inflation rate [B,C]
Program Loan	5155	0.14	0.35	0	1	=1 if Program Loan
SIL	5155	0.54	0.50	0	1	=1 if Specific Investment Loan
East Asia-Pacific	5155	0.16	0.36	0	1	=1 if East Asia-Pacific region
Europe & Central Asia	5155	0.13	0.33	0	1	=1 if Europe & Central Asia region
Latin America & Caribbean	5155	0.20	0.40	0	1	=1 if Latin America & Caribbean region
Middle East & North Africa	5155	0.09	0.28	0	1	=1 if Middle East & North Africa region
South Asia	5155	0.13	0.33	0	1	=1 if South Asia region
Sub-Saharan Africa	5155	0.31	0.46	0	1	=1 if Sub-Saharan Africa region
log US Aid	4845	4.07	1.50	0	8.58	log of (U.S. bilateral aid + 1) in 2005 USD (millions) [B,C]
log G7 Aid	4845	6.17	1.32	0.98	8.93	log of (G7 bilateral aid + 1) in 2005 USD (millions) [B,C]
US military aid	5140	0.41	0.49	0	1	=1 if U.S. military aid>2 million (2012 USD) [B]
US UN voting	5154	0.42	0.17	0.00	0.85	UN voting alignment with U.S. on important votes [B]
G7 UN voting	5154	0.63	0.15	0.26	0.95	UN voting alignment with G7 on important votes [B]
UNSC (*ICR*)	5155	0.08	0.28	0	1	=1 if nonpermanent UNSC member [B]
WBEB (*ICR*)	5155	0.33	0.47	0	1	=1 if Executive Board member in last 3 years [B]

[A] Missing values imputed using Sustainability, Supervision, Institutional Development Impact, and Risk to Development Outcome ratings.
[B] At time of most recent ICR.
[C] Three year moving average (*t*-2 to *t*).

Table 3. Baseline Hazard Rate for PPAR

	(1)	(2)	(3)	(4)	(5)
Outcome (*ICR*)	1.476***	1.609***	0.933		1.592***
	(4.33)	(4.62)	(-0.89)		(4.42)
Unsatisfactory				2.405**	
				(2.35)	
Moderately Unsatisfactory				2.955***	
				(2.85)	
Moderately Satisfactory				3.659***	
				(3.40)	
Satisfactory				3.936***	
				(3.77)	
Highly Satisfactory				4.638***	
				(4.13)	
IDA	1.080	1.132	0.788**	1.193	1.197
	(0.76)	(1.02)	(-2.30)	(1.52)	(1.54)
ICR quality	0.497***	0.464***	1.070	0.551***	0.559***
	(-7.82)	(-6.42)	(0.62)	(-5.76)	(-5.64)
log World Bank debt	1.134*	1.157*	1.038	1.111	1.107
	(1.72)	(1.77)	(0.56)	(1.33)	(1.29)
log Project size	1.140***	1.178***	1.029	1.156***	1.149***
	(3.03)	(3.28)	(0.81)	(3.00)	(2.89)
log of # WB projects	1.137**	1.157*	1.051	1.141*	1.148*
	(2.01)	(1.81)	(0.88)	(1.78)	(1.84)
June ICR	0.854**	0.822**	0.929	0.828**	0.828**
	(-2.07)	(-2.05)	(-0.83)	(-2.03)	(-2.04)
Tourism	1.161***	1.176**	1.124**	1.171***	1.171***
	(2.77)	(2.55)	(2.05)	(2.70)	(2.67)
Years in office	1.011**	1.014**	0.998	1.011*	1.011*
	(2.01)	(1.99)	(-0.53)	(1.82)	(1.85)
Freedom House	0.879***	0.862***	1.012	0.893***	0.889***
	(-3.81)	(-3.66)	(0.35)	(-2.72)	(-2.84)
log Population	0.874*	0.859*	0.899	0.851**	0.855*
	(-1.84)	(-1.82)	(-1.44)	(-2.00)	(-1.93)
log GDP PC	0.709***	0.660***	1.025	0.698***	0.699***
	(-3.71)	(-3.58)	(0.28)	(-3.52)	(-3.48)
GDP growth	1.017	1.019	1.021*	1.009	1.011
	(1.36)	(1.34)	(1.91)	(0.64)	(0.84)
Inflation	0.997	0.989	1.024	1.168	1.127
	(-0.04)	(-0.12)	(0.69)	(0.93)	(0.63)
Program Loan	2.035***	2.006***	0.987	1.686***	1.710***
	(6.14)	(4.68)	(-0.12)	(4.20)	(4.33)
× *Inflation*	1.464***	2.364**	1.214**	1.299	1.358
	(3.19)	(2.11)	(2.08)	(1.45)	(1.54)

SIL	1.171**	1.177*	1.094	1.053	1.053
	(2.18)	(1.90)	(1.47)	(0.57)	(0.57)
East Asia-Pacific	0.888	0.909	0.716**	0.892	0.899
	(-0.97)	(-0.66)	(-2.47)	(-0.83)	(-0.77)
Europe & Central Asia	1.747***	1.950***	0.877	1.601**	1.609**
	(3.13)	(3.09)	(-0.95)	(2.43)	(2.47)
Latin America & Caribbean	1.019	1.088	0.864	1.136	1.142
	(0.12)	(0.43)	(-1.00)	(0.71)	(0.74)
Middle East & North Africa	1.097	1.216	0.761**	1.043	1.039
	(0.50)	(0.92)	(-2.03)	(0.22)	(0.20)
South Asia	0.731	0.729	1.156	0.895	0.880
	(-1.46)	(-1.34)	(0.95)	(-0.47)	(-0.52)
Observations	5155	5047	1371	4282	4282

z-statistics in parentheses based on country-clustered standard errors. All specifications include unreported evaluation year dummies. Hazard models use a Weibull regression; all coefficients reported as hazard or odds ratios.

(1) Hazard model with dichotomous ICR Outcome rating, full sample.

(2) Logit PAR selection model (probability of being selected for PPAR by September 30, 2013); some observations drop due to lack of variation by year.

(3) Hazard model with dichotomous ICR Outcome rating, uncensored sample.

(4) Hazard model with dummy variables reflecting 6-point ICR Outcome rating; omitted category is "Unsatisfactory." Sample starts in 1995 with the introduction of 6-point scale.

(5) Hazard model with dichotomous ICR Outcome rating, =1 if rating is "Moderately Satisfactory" or above. Sample constrained to match (4).

Table 4. Hazard Ratios for PPAR -- Political Economy variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
log US Aid	1.075**						
	(2.13)						
log G7 Aid		1.181**					
		(2.31)					
US military aid			0.998				
			(-0.02)				
US UN voting				1.139			
				(0.36)			
G7 UN voting					0.927		
					(-0.19)		
UNSC (*ICR*)						0.970	
						(-0.21)	
WBEB (*ICR*)							0.976
							(-0.21)
Observations	4845	4845	5140	5154	5154	5155	5155

z-statistics in parentheses based on country-clustered standard errors. All specifications include baseline variables from Table 3, Column 1. Number of observations vary due to data availability. Estimates from hazard function with Weibull distribution reported as hazard ratios.

Table 5. Role of rating changes. PPAR hazard ratios for projects with completed PPARs.

	(1)	(2)	(3)	(4)
Outcome (*ICR*)	0.933	0.945	0.950	0.952
	(-0.89)	(-0.69)	(-0.61)	(-0.58)
Downgrade		0.705***	0.792**	0.771**
		(-3.81)	(-2.08)	(-2.37)
Upgrade		0.749	0.640*	0.654*
		(-1.54)	(-1.86)	(-1.83)
WBEB (*ICR*)			0.887	0.870*
			(-1.45)	(-1.76)
× *Downgrade*			0.702**	0.713**
			(-2.13)	(-1.99)
× *Upgrade*			1.760**	1.663*
			(2.02)	(1.95)
UNSC (*ICR*)				1.012
				(0.08)
UNSC (*PPAR*)				0.645***
				(-3.49)
IDA	0.788**	0.793**	0.790**	0.751***
	(-2.30)	(-2.16)	(-2.13)	(-2.68)
ICR quality	1.070	1.032	1.030	1.038
	(0.62)	(0.30)	(0.30)	(0.40)
log World Bank debt	1.038	1.029	1.025	1.041
	(0.56)	(0.43)	(0.40)	(0.65)
log Project size	1.029	1.033	1.033	1.036
	(0.81)	(0.90)	(0.87)	(0.96)
log # WB projects	1.051	1.046	1.040	1.039
	(0.88)	(0.80)	(0.74)	(0.71)
June ICR	0.929	0.902	0.901	0.908
	(-0.83)	(-1.11)	(-1.14)	(-1.11)
Tourism	1.124**	1.134**	1.145**	1.120**
	(2.05)	(2.26)	(2.54)	(2.13)
Years in office	0.998	0.997	0.996	0.998
	(-0.53)	(-0.75)	(-0.94)	(-0.48)
Freedom House	1.012	1.016	1.009	1.015
	(0.35)	(0.48)	(0.26)	(0.47)
log Population	0.899	0.891	0.907	0.919
	(-1.44)	(-1.59)	(-1.38)	(-1.22)
log GDP PC	1.025	1.019	1.018	1.047
	(0.28)	(0.21)	(0.21)	(0.53)
GDP growth	1.021*	1.017	1.018	1.022**
	(1.91)	(1.57)	(1.57)	(2.00)
Inflation	1.024	1.031	1.031	1.056
	(0.69)	(0.85)	(0.88)	(1.45)
Program Loan	0.987	0.983	0.991	1.012
	(-0.12)	(-0.17)	(-0.09)	(0.11)

× *Inflation*	1.214**	1.181*	1.173*	1.152
	(2.08)	(1.80)	(1.71)	(1.45)
SIL	1.094	1.083	1.073	1.084
	(1.47)	(1.23)	(1.08)	(1.18)
East Asia-Pacific	0.716**	0.721**	0.733**	0.674***
	(-2.47)	(-2.34)	(-2.27)	(-3.05)
Europe & Central Asia	0.877	0.866	0.828	0.788*
	(-0.95)	(-1.05)	(-1.32)	(-1.76)
Latin America & Caribbean	0.864	0.869	0.889	0.884
	(-1.00)	(-0.96)	(-0.81)	(-0.90)
Middle East & North Africa	0.761**	0.771*	0.758**	0.692***
	(-2.03)	(-1.91)	(-2.04)	(-2.89)
South Asia	1.156	1.206	1.229	1.186
	(0.95)	(1.32)	(1.55)	(1.27)
Observations	1371	1371	1371	1371

z-statistics in parentheses based on country-clustered standard errors. All specifications include unreported evaluation year (ICR) dummies. Weibull regression with coefficients reported for log relative-hazard form.

Table 6. PPAR ratings and non-permanent UNSC membership.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ICR	2.050***	0.695***	0.693***		2.426***	0.774***	0.676***	
	(23.91)	(34.38)	(34.04)		(17.77)	(28.71)	(29.36)	
*ICR*2 (Unsatisfactory)				0.464				0.806***
				(1.01)				(2.75)
*ICR*3 (Moderately Unsatisfactory)				1.137**				1.350***
				(2.41)				(4.44)
*ICR*4 (Moderately Satisfactory)				2.034***				2.302***
				(4.44)				(7.97)
*ICR*5 (Satisfactory)				2.575***				2.869***
				(5.72)				(10.74)
*ICR*6 (Highly Satisfactory)				3.533***				3.684***
				(7.82)				(11.99)
UNSC (*PAR*)	0.312**	0.0978**	0.0670**	0.223***	0.333*	0.0945**	0.0643**	0.232***
	(2.25)	(2.48)	(2.42)	(3.53)	(1.89)	(2.15)	(2.27)	(2.99)
UNSC (*ICR*)	-0.0437	-0.0153	-0.0103	-0.0602	0.0464	0.0147	-0.00290	-0.0818
	(-0.26)	(-0.26)	(-0.26)	(-0.60)	(0.25)	(0.25)	(-0.10)	(-1.18)
UNSC (*approval*)	0.0404	0.0138	0.00877	-0.0454	0.128	0.0393	0.0258	-0.0379
	(0.36)	(0.36)	(0.34)	(-0.84)	(0.77)	(0.80)	(0.85)	(-0.62)
Observations	1500	1500	1500	1500	1219	1219	1329	1329

z/*t*-statistics in parentheses based on country-clustered standard errors.
Probit on binary PPAR rating: (1) Basic specification, (5) Full specification
Marginal effects for Probit: (2) Basic specification, (6) Full specification
OLS on binary PPAR rating: (3) Basic specification, (7) Full specification
OLS on 6-point PPAR rating: (4) Basic specification, (8) Full specification
Marginal effects for ICR rating calculated with other variables set at sample mean. Marginal effects for each UNSC variable calculated with other UNSC variables at zero and all other variables at sample mean.
* p<.1, ** p<.05, *** p<.01

Table 7. Interaction of UNSC membership and World Bank Debt.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ICR	2.063***	0.698***	0.694***		2.431***	0.775***	0.671***	
	(24.23)	(35.00)	(35.18)		(17.27)	(27.67)	(29.16)	
*ICR*2 (Unsatisfactory)				0.467				0.777***
				(1.00)				(2.80)
*ICR*3 (Moderately Unsatisfactory)				1.170**				1.336***
				(2.42)				(4.59)
*ICR*4 (Moderately Satisfactory)				2.047***				2.284***
				(4.37)				(8.34)
*ICR*5 (Satisfactory)				2.578***				2.836***
				(5.60)				(11.28)
*ICR*6 (Highly Satisfactory)				3.531***				3.654***
				(7.66)				(12.51)
UNSC (*PPAR*) × *Low Debt*	-0.0992	-0.0357	-0.0251	-0.0206	-0.557	-0.219	-0.0837	-0.188
	(-0.29)	(-0.28)	(-0.30)	(-0.12)	(-1.29)	(-1.33)	(-1.15)	(-1.00)
UNSC (*PPAR*) × *Medium Debt*	0.310	0.0984*	0.0670	0.250**	0.220	0.0592	0.0613	0.271**
	(1.51)	(1.66)	(1.63)	(2.53)	(0.93)	(1.01)	(1.59)	(2.36)
UNSC (*PPAR*) × *High Debt*	0.482***	0.135***	0.0904***	0.214**	0.753***	0.152***	0.119***	0.298**
	(3.90)	(3.70)	(3.46)	(2.31)	(4.16)	(2.63)	(3.65)	(2.50)
Low Debt	-0.0115	-0.0202	-0.00291	-0.0889	-0.571*	0.0231	-0.128**	-0.0932
	(-0.10)	(-0.35)	(-0.11)	(-1.14)	(-1.72)	(0.39)	(-1.98)	(-0.63)
High Debt	0.0831	0.00808	0.0198	0.0857	0.0241	0.0294	0.0328	0.116
	(0.62)	(0.21)	(0.63)	(1.09)	(0.07)	(0.57)	(0.42)	(0.55)

UNSC (*ICR*)	-0.0576	-0.00404	-0.0139	-0.0808	0.0748	-0.199	-0.00134	-0.0742
	(-0.35)	(-0.10)	(-0.36)	(-0.87)	(0.38)	(-1.61)	(-0.05)	(-1.04)
UNSC (*approval*)	0.0235	0.0281	0.00489	-0.0816	0.0958	0.00679	0.0226	-0.0438
	(0.21)	(0.62)	(0.19)	(-1.51)	(0.56)	(0.07)	(0.74)	(-0.70)
Observations	1451	1451	1451	1451	1219	1219	1329	1329

z/*t*-statistics in parentheses based on country-clustered standard errors.
Probit on binary PPAR rating: (1) Basic specification, (5) Full specification
Marginal effects for Probit: (2) Basic specification, (6) Full specification
OLS on binary PPAR rating: (3) Basic specification, (7) Full specification
OLS on 6-point PPAR rating: (4) Basic specification, (8) Full specification
Marginal effects for ICR rating calculated with other variables set at sample mean. Marginal effects for each UNSC variable calculated with other UNSC / debt variables at zero and all other variables at sample mean.
* p<.1, ** p<.05, *** p<.01

Appendix: Data Sources

Outcome (ICR)	IEG (2014)
IDA	World Bank (2012B)
ICR quality	IEG (2014)
log World Bank debt	World Bank (2014)
log Project size	World Bank (2014)
log # WB projects	IEG (2014)
June ICR	IEG (2014)
Tourism	World Bank (2014)
Years in office	Beck et al. (2001)
Freedom House	Freedom House (2012)
log Population	World Bank (2014)
log GDP PC	World Bank (2014)
GDP growth	World Bank (2014)
Inflation	World Bank (2014)
Program Loan	World Bank (2012B)
SIL	World Bank (2012B)
East Asia-Pacific	World Bank (2012B)
Europe & Central Asia	World Bank (2012B)
Latin America & Caribbean	World Bank (2012B)
Middle East & North Africa	World Bank (2012B)
South Asia	World Bank (2012B)
Sub-Saharan Africa	World Bank (2012B)
log US Aid	OECD (2015)
log G7 Aid	OECD (2015)
US military aid	USAID (2014)
US UN voting	Strezhnev and Voeten (2012)
G7 UN voting	Strezhnev and Voeten (2012)
UNSC (*ICR*)	United Nations (2014)
WBEB (*ICR*)	Kaja and Werker (2010)