RE-379

EVALUABILITY REVIEW OF BANK PROJECTS 2009

Office of Evaluation and Oversight, OVE

Inter-American Development Bank
Washington, D.C.
October 2010

For Official Use Only

TABLE OF CONTENTS

ABBREVIATIONS

I. INTRODUCTION 1

II. EVALUABILITY OF PROJECTS 2

III. MANAGEMENT'S REVIEW PROCESS 12

IV. REVIEW OF THE DEVELOPMENT EFFECTIVENESS MATRIX (DEM) 20

V. MANAGEMENT'S RESPONSE 25

VI. CONCLUSIONS AND RECOMMENDATIONS 28

ENDNOTES

ANNEXES

ANNEX I: DISTRIBUTION OF WEIGHTS

http://idbdocs.iadb.org/WSDocs/getdocument.aspx?docnum=35409100

ANNEX II: EVALUABILITY NOTES

http://idbdocs.iadb.org/WSDocs/getdocument.aspx?docnum=35409080

ANNEX III: EXTENDED TABULATIONS

http://idbdocs.iadb.org/WSDocs/getdocument.aspx?docnum= 35409106

ABBREVIATIONS

CCLIP	Conditional Credit Line for Investment Projects
CRM	Credit Review Meeting
DEF	Development Effectiveness Framework
DEM	Development Effectiveness Matrix
EME	Emergency Loan
ERM	Eligibility and Risk Review Meeting
ERR	Economic Rate of Return
ESG	Environmental and Safeguard Unit
FIN	Finance Department
LEG	Legal Department
MIF	Multilateral Investment Fund
NLF	New Lending Framework
NPC	New Project Cycle
NSG	Non-Sovereign Guarantee
OMJ	Opportunities for the Majority
OPC	Operations and Policy Committee
OVE	Office of Evaluation and Oversight
PBL	Policy-Based Loan
PLB	Programmatic PBL
POD	Project Operation Document
PP	Project Profile
QRR	Quality and Risk Review
RES	Research Department
RMG	Office of Risk Management
SG	Sovereign Guarantee
SPD	Office of Strategic Planning and Development Effectiveness
SPD	Office of Strategic Planning and Development
SUP	Supplemental Financing Operations
VPC	Vice-Presidency for Countries
VPP	Vice-Presidency for Private Sector
VPS	Vice-Presidency for Sectors

I. INTRODUCTION

1.1 Evaluability is a simple concept: *it is the ability of an intervention to demonstrate in measurable terms the results it intends to deliver*. Measuring results is necessary for a project or program to be managed by results, and it is also necessary for institutional learning regarding development effectiveness.

1.2 In 2001 the Office of Evaluation and Oversight (OVE) developed a method for assessing evaluability and applied that tool to all of the projects approved that year. A similar exercise using the same tool was completed in 2005. In designing its evaluability instrument, OVE reviewed and adapted existing methodologies for assessing evaluability[1]. The method used by OVE is an assessment based on independent peer reviews, which relies on the expert findings of reviewers to assess proposals in different dimensions of evaluability[2]. OVE's protocol is similar to methodologies applied in the qualitative review of funding proposals, as well as those applied more broadly in the quality review of projects' design and analytical work[3].

1.3 When designing the evaluability instrument, OVE adhered to overarching principles associated with the review of analytical work[4]. These include *Accountability, Quality, Inter-subjectivity, and Independence.* To ensure *Accountability* one person is responsible for the exercise implementation and for the final outcome, including the quality of the peer reviews, the documentation of findings, and the reporting of findings. *Quality* is assured by employing a two-step assessment in which an in-depth assessment of loan proposals is complemented with a quantitative scoring instrument based on a multinomial rating scale. The use of *Inter-subjectivity* allows for the richness of individual expert reviews, while at the same time minimizing reviewer bias by employing a panel of reviewers. Lastly, *Independence* of the exercise is assured by utilizing reviewers who were not involved in project preparation, and by preserving their anonymity.

1.4 The evaluability analysis includes a set of <u>formal</u> dimensions, such as the identification of appropriate output and outcome indicators, baselines, and an adequate monitoring and evaluation strategy. It also includes <u>substantive</u> dimensions—characteristics that define the substance of the intervention itself, which are necessary for the proper identification of the formal evaluation dimensions. These include the identification of a problem through a diagnosis, the identification of what is hoped to be achieved through the definition of objectives, the identification of how this is to be achieved through its intervention logic, and the identification of risks that could attenuate effectiveness.

1.5 The review process includes the following steps. First, OVE's Deputy Director was assigned responsibility for the entire evaluability review exercise, aided by an Evaluability panel comprised of three staff members. Second, an evaluability team was assigned to each project reviewed. This team consisted of a team leader,

and a minimum of three other members. The team assessed the project document information (loan document and annexes) and met for discussion. A draft of findings in each of the evaluability dimensions was prepared in the form of an *Evaluability Note (Note)*. Third, the findings were discussed with the evaluability panel. This ensured that assessments were done according to the established evaluability criteria. Fourth, if the panel decided that changes to the *Note* were needed, these were then made by the team. Finally, and at a subsequent meeting, the panel and the team rated the project in each of the evaluability dimensions. The findings of the *Note* form the basis for these ratings. These ratings are recorded. The ratings of projects are on a four-point scale, with two ratings indicating inadequate content (1 and 2) and the two indicating adequate content (3 and 4). In each case, at least seven reviewers were involved in the process including three members in each project teams, and four members of the core panel[5].

II. EVALUABILITY OF PROJECTS

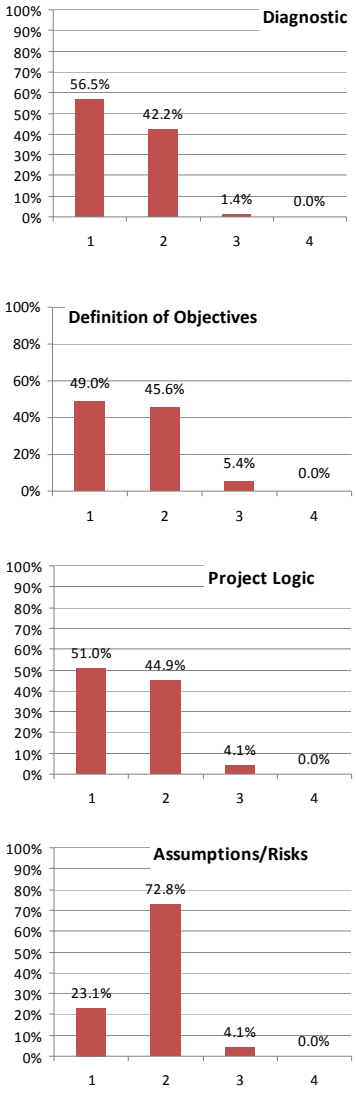
2.1 In 2009 OVE assessed the evaluability of all projects by examining the loan materials circulated to the Board[6]. The percentage or projects achieving an adequate or better rating—that is, a rating of 3 or 4 on a 4-point scale—was low for all evaluability dimensions, ranging from 1.4% in the case of diagnostic, to 45% in the case of output indicators. Results on each of the nine evaluability dimensions are summarized below; selected examples are presented in the footnotes corresponding to each finding, and a full description of the problems is found in the 147 *Notes*, annexed to this report. Results for the substantive dimensions of Evaluability—Diagnostic, Objectives, Logic and Risks—are presented first. These are followed by results for the formal dimensions of Evaluability—Output and Outcome indicators and baselines, and Monitoring and Evaluation.

2.2 **Diagnostics**. Diagnostics should provide the information required to understand the nature and scope of the development problems being addressed, and to understand the role that the proposed intervention will have in addressing these problems. In 2009 projects did not do this. Only 2 of the 147 projects (1.4%) had an adequate rating for diagnostic.

2.3 The review found that LDs did not provide an adequate identification of key development challenges, and did not present the data and analysis necessary to properly dimension the magnitude of challenges[7]. They also did not provide the data and analysis required in order to link the proposed solutions to the identified development problem[8]. In other instances diagnostics information is provided, but not the relevant information for the underlying causes of the development issues identified[9].

2.4 In the case of private sector operations, this is seen in the lack of data on market functioning and failures, as well as positive and negative externalities. For instance, in the case of the Bank's *"Green Bank" Initiatives*—operations which finance the construction of environmentally friendly headquarters for Banks—the LD does not discuss the underlying market failures or other reasons why Banks and other institutions had not adopted green buildings, and why they are doing so now[10].

Figure 2.1: Evaluability Ratings: Substantive Dimensions



2.5 LD diagnostics did not adequately identify benefits and beneficiaries. In some cases this is due to an ambiguous definition of who the project would target[11]. In the case of some Opportunity for the Majority (OMJ) operations, this was seen with respect to the inadequate definition of "majority"[12].

2.6 In order to properly justify selected interventions as a solution to identified problems, diagnostics should provide evidence of the effectiveness of proposed interventions, particularly in areas where the Bank can draw on its own experience and analysis. LDs did not do this. Even in contexts where information is available and was collected by the countries, this was not utilized as a basis for the proposed intervention components[13]. In cases where the Bank is in second, third, or even higher operations in the same sector and the same country, there is also little identification of lessons learned and applied, at least in terms of focusing the effectiveness of programs[14].

2.7 **Definition of Objectives**. The definition of objectives makes up the programmatic intent of the Bank. Objectives should identify what is expected to be achieved, in terms of welfare improvements and development results. Objectives have to be specific, measurable, attainable, relevant and time-bound. In 2009, projects did not present evaluable objectives. Only 8 of 147 projects (5.4%) had an adequate identification of objectives.

2.8 LDs identified objectives centered on activities to be achieved, rather than on development results or welfare improvements[15]. LDs also used qualifiers in their

objectives that dilute the clarity of intent and accountability (e.g. "to help the country", "to support the sector", "to contribute to reform")[16]. LDs contained multiple and often conflicting statements of intent[17], or in some cases—particularly in the case of Non-Sovereign Guarantee (NSG) operations—projects contained no explicit statement of intent.

2.9 **Project Logic** The project's logic is the cornerstone of the intervention. It provides the rational and empirical link between problem, objective and solution. It also provides the link between activities, outputs and outcomes. In 2009 project logic adequate in only 6 of 147 projects (4.1%).

2.10 LDs did not provide a plausible justification for the interventions or an adequate explanation for their *raison d'être*. In such cases, LDs did not identify the relationship between problems, objectives and solutions, or did not demonstrate that the problems outlined in the diagnosis could be solved through the proposed intervention[18]. In other cases the information provided is insufficient to provide a minimum understanding of how problems can be addressed or why, or to understand if the magnitude of the need is consistent with the specified amount of funding[19]. The link between problem and solution is particularly weak in projects which identify the financial crisis as a motivation for the proposed financing. In these cases, which include both emergency lending and investment lending, the role of the financial crisis as either the cause of the need or as an aggravating factor, is not adequately justified[20].

2.11 LDs provided a justification for interventions, but this justification was inconsistent with the data and evidence provided in the diagnostic. This produces an inconsistency between development solutions proposed and the evidence provided in the LD. In other instances, this inconsistency was simply due to lack of a diagnostic[21].

2.12 The causal chain between inputs→outputs→outcomes was not clearly established in the LD.[22] In such cases the LD and the Result Framework did not adequately present or articulate the different elements of project design. They presented potentially conflicting components[23]; and internal inconsistencies.[24] In many instances the project intervention models were not made explicit, or in some instances the project logic that is made explicit does not deal with the causes of development challenges, but rather with their proximate manifestations. In some instances the corresponding sequence is chronologically inconsistent.[25]

2.13 Operations defined as demonstration experiences or pilot projects lacked an adequate explanation of the model to be tested. In such cases, despite being presented as vehicles for learning, they did not incorporate the means for ascertaining results or for ascertaining the channels by which results were produced[26]. The models on which project design is predicated were not explicitly presented or explained[27]. Project documents state that their design have been

based on best practices or results of earlier phases or prior operations, but do not state, identify or document what these lessons were[28].

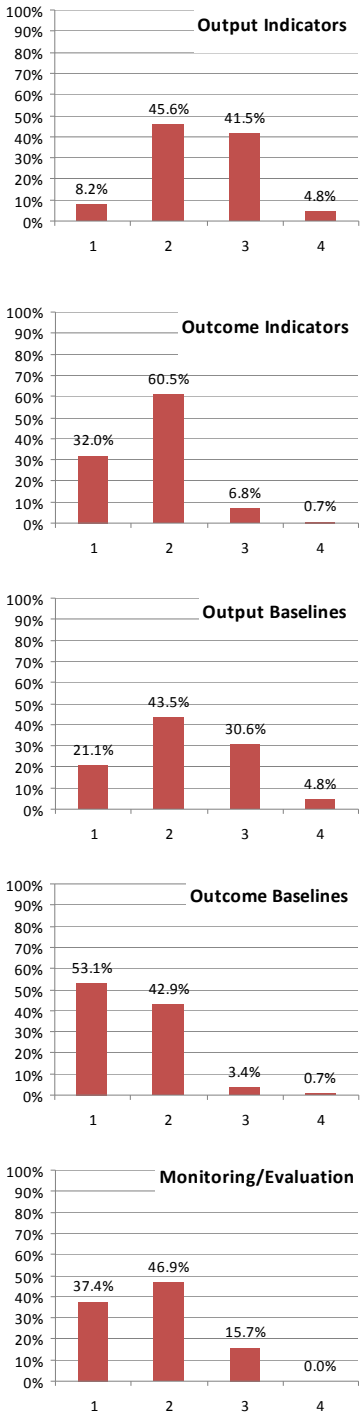
2.14 In the case of NSG projects, the LD does not demonstrate the Bank's additionality. In the private sector, projects often claim financial and non-financial additionalities in order to justify their interventions. However, in most cases, such additionalities are unsupported by evidence and are not consistent with the proposed interventions[29].

2.15 **Assumptions and Risks**. Risks are important because they identify the major factors that can impact project performance and achievement of results. A mitigation measure is the identification of activities that can either lessen the likelihood of a risk, or can lessen the impact of said factors on the project's performance. Only 6 of 147 projects (4.1%) had adequate identification of risks and mitigation measures.

2.16 Most LDs provided a discussion of fiduciary and environmental risks, and in some cases execution risks. However, in many cases these risks are not adequately identified[30]. In the case of NSG projects risks associated with the absence of a B lender are also not adequately identified[31]. Other risks, such as sector specific risks, institutional, and reputation risks are not identified[32].

2.17 LDs identified elements of project components as project risks; since these elements are being treated by the program, identifying them also as risks is not meaningful[33]. In other cases LDs were erroneously identified as risks events that had already materialized[34]. Lastly, mitigation measures were inadequate and tracking mechanisms were absent.[35]

Figure 2.2: Evaluability Ratings: Formal Dimensions



2.18 **Output Indicators.** Output indicators measure the delivery of project goods and services, or in the case of policy reform, measure activities involved with the

reform process (e.g. legislative or normative changes). Sixty-eight of 147 projects (47%) had an adequate definition of outputs. This was the highest percentage of adequate of any of the evaluability ratings in 2009.

2.19 Despite the higher evaluability ratings for output indicators, issues remain. Projects identified the delivery of goods and services, but did not measure aspects of quality[36]. This is the case, for example, of projects that offer training programs in which the indicator is only the number of people trained, regardless of the quality of the training.[37] Output indicators were also missing for some portion of the project's components[38]; or they were not sufficiently clear to understand what was being measured[39]. The assessment also found that projects identify indicators, but do not set targets and/or milestones or do so in an unclear manner[40].

2.20 Lastly, the assessment also identified a problem with presentation, in which content required to understand what is to be delivered and when is in some instances present in different parts of the project (e.g. the loan document, the results matrix and the annexes). This is particularly important in the cases where inconsistency between parts of the document can be observed, either in the definition of the indicator or in the milestones and targets[41]. This problem is present in the case of both output indicators and outcome indicators.

2.21 **Outcome Indicators**. Outcome indicators measure the welfare and development impact of the intervention. These are the indicators that measure the program impact on beneficiaries. Eleven of 147 projects (8%) contained an adequate identification of outcome indicators.

2.22 The main findings with respect to outcome indicators were similar to those documented in prior evaluability reviews. Outcome indicators focused on outputs and activities, rather than on results. Outcome indicators did not properly track project development objectives, or outcome indicators did not span the entirety of purported benefits. The evaluation also found that outcome indicators did not correspond to the beneficiary population, went beyond the scope of the project or did not correspond to the time period of the project's intervention[42].

2.23 **Output Baselines.** Output indicator baselines identify the stock of goods and services ex ante. In cases where projects are introducing new concepts or defining new deliverables, baselines are taken as being the absence of the proposed activities. In cases where baselines are building upon an existing stock of services or goods, baselines refer to that existing stock. Fifty-two of 147 projects (35%) had an adequate definition of baselines for their outputs.

2.24 The evaluability problems detected in output baselines include that (i) projects simply lacked the output indicators baselines[43], without any reference to whether such measures will be collected, and (ii) projects which were building on an existing stock of outputs identified baselines as zero, rather than the existing stock[44].

2.25 **Outcome Baselines.** Outcome baselines define the problem situation ex ante. Without baselines it is not possible to adequately dimension the size of the identified problem. Baselines are also essential for assessment of project results in almost all circumstances. Only six of 147 projects (4%) properly identified baselines.

2.26 Findings regarding outcome baselines include that (i) Baselines were not identified[45], or were identified for populations which were not the intended beneficiaries; (ii) Baselines were outdated, and no longer represented the current situation with respect to project results[46]; and (iii) In some cases there is evidence in the document suggesting that the baseline was not properly computed or reported, or different baselines are reported in distinct parts of the document[47].

2.27 **Monitoring and Evaluation.** Monitoring and evaluation frameworks should identify the resources, responsibilities, data collection, and method required to track and estimate project results. Out of 147 projects, only 23 had an adequate monitoring and evaluation systems designed.

2.28 Findings regarding monitoring and evaluation were: (i) projects lacked any identification of responsibilities, data collection, financing[48], or method/questions. In short, many projects do not have a complete evaluation framework; (ii) projects identified a monitoring system, but the system did not include an impact evaluation of project results[49]; or the system did not focus on the intervention's innovations; (iii) projects used imprecise, and in some instances incorrect evaluation concepts and techniques[50].

2.29 Table 2.1 details evaluability results for the Bank's instruments. It also summarizes the results of the most common instruments within Sovereign Guarantee (SG) operations and for the Bank's OMJ operations. Results are presented as the percentage of projects which score adequate or above in each of the evaluability dimensions.

2.30 The first result seen is that all instruments performed poorly. SG operations did better than NSG operations in all formal categories, except in outcome indicators, where they performed comparably. This is a reversal of the pattern observed in 2005, when NSG operations were clearly more evaluable than SG ones. In substantive categories almost no project was rated satisfactory[51]. In risks NSG operations did somewhat better, and in the definition of objectives. However, the latter is driven entirely by the better performance of OMJ projects. OMJ projects performance in substantive dimensions was comparable to or better than SG operations. This bimodal distribution within NSG operations reflects a better identification of market failures and externalities for OMJ operations, and a definition of objectives that went beyond private returns for the borrower.

2.31 With respect to a monitoring framework, again there were no NSG projects with an adequate rating. Finally, few NSG operations properly calculated ERRs. In most cases the concepts used to measure economic returns did not adequately identify externalities. In general, NSG operations did not identify the indicators, data sources, and baseline information that would allow for tracking project performance.

Table 2.1: Percentage of projects with satisfactory of better scores, by instrument (top), country group and Bank unit (bottom)

Instrument		Substantive Dimensions				Formal Dimensions				
	Count	Diagnostic	Objectives	Logic	Risks	Output Indicators	Outcome Indicators	Output Baselines	Outcome Baselines	Monitoring Evaluation
Sovereign Guarantee	**119**	**0.8%**	**4.2%**	**4.2%**	**3.4%**	**52.9%**	**7.6%**	**39.5%**	**5.0%**	**19.3%**
Multi-phase (PFM)	12	0.0%	0.0%	8.3%	8.3%	75.0%	16.7%	66.7%	8.3%	41.7%
Global Multiple Works (GOM)	6	0.0%	0.0%	16.7%	16.7%	50.0%	16.7%	33.3%	16.7%	16.7%
Global Credit (GCR)	6	0.0%	16.7%	0.0%	0.0%	83.3%	0.0%	66.7%	0.0%	0.0%
PBL (*)	16	0.0%	0.0%	6.3%	0.0%	68.8%	0.0%	56.3%	0.0%	6.3%
Other (**)	8	0.0%	12.5%	12.5%	0.0%	25.0%	12.5%	0.0%	12.5%	50.0%
Specific Investment (ESP)	43	2.3%	7.0%	2.3%	2.3%	44.2%	7.0%	32.6%	2.3%	23.3%
of which CCLIP (CLP)	18	0.0%	0.0%	0.0%	5.6%	50.0%	11.1%	38.9%	11.1%	11.1%
Sup. Financing (SUP)	6	0.0%	0.0%	0.0%	0.0%	50.0%	0.0%	33.3%	0.0%	0.0%
Financial Emergency (EME)	4	0.0%	0.0%	0.0%	0.0%	50.0%	0.0%	25.0%	0.0%	0.0%
Non-sovereign Guarantee	**28**	**3.6%**	**10.7%**	**3.6%**	**7.1%**	**17.9%**	**7.1%**	**17.9%**	**0.0%**	**0.0%**
of which OMJ	*7*	14.3%	42.9%	14.3%	14.3%	*14.3%*	*28.6%*	*14.3%*	*0.0%*	*0.0%*
of which SCF	*21*	0.0%	0.0%	0.0%	4.8%	*19.1%*	*0.0%*	*19.1%*	*0.0%*	*0.0%*

(*) includes 15 programmatic and two traditional PBL

(**) includes ERF, FAB,FAC,IGR,INO,PDL, and TCR, of which there were only one or two operations approved

		Substantive Dimensions				Formal Dimensions				
	Count	Diagnostic	Objectives	Logic	Risks	Output Indicators	Outcome Indicators	Output Baselines	Outcome Baselines	Monitoring Evaluation
Region										
Andean (CAN)	33	*3.0%*	*6.1%*	*9.1%*	*9.1%*	*60.6%*	*9.1%*	*48.5%*	*3.0%*	*12.1%*
Southern Cone (CSC)	49	*2.0%*	*6.1%*	*4.1%*	*4.1%*	*55.1%*	*10.2%*	*40.8%*	*4.1%*	*20.4%*
CA, ME, DR, PN (CID)	45	*0.0%*	*4.4%*	*2.2%*	*2.2%*	*35.6%*	*4.4%*	*28.9%*	*4.4%*	*11.1%*
Caribbean (CCB)	19	*0.0%*	*5.3%*	*0.0%*	*0.0%*	*26.3%*	*5.3%*	*15.8%*	*5.3%*	*21.1%*
Bank Unit (*)										
Social (SCL)	*17*	*5.9%*	*0.0%*	*11.8%*	*5.9%*	*70.6%*	*17.7%*	*64.7%*	*5.9%*	*58.8%*
Integration, Trade (INT)	*2*	*0.0%*	*0.0%*	*0.0%*	*0.0%*	*50.0%*	*50.0%*	*50.0%*	*50.0%*	*0.0%*
Infrastructure, Env. (INE)	*57*	*0.0%*	*5.3%*	*5.3%*	*3.5%*	*43.9%*	*7.0%*	*35.1%*	*5.3%*	*15.8%*
Institutional, Finance (ICF)	*43*	*0.0%*	*4.7%*	*0.0%*	*2.3%*	*58.1%*	*2.3%*	*34.9%*	*2.3%*	*9.3%*

(*) VPP is excluded, as the responsible unit uniquely identifies the sector, which is already reported above (see NSG and OMJ above)

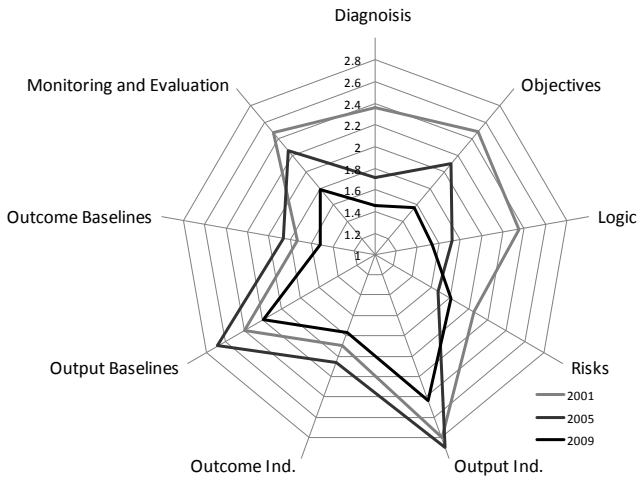
2.32 The evaluability performance of SG operations was also different across lending modality, with modalities such as the financial emergency lending and the supplemental financing lending performing substantially worse. Supplemental financing operations and emergency financing operations were the only lending modalities that did not have any project with an adequate rating on any of the substantive or formal dimensions, except those associated with outputs. It should be noted that in the case of supplemental financing, the projects were evaluated with respect to the causes of the cost overrun, and the identification of the remedy[52]. In 2009 programmatic PBLs had replaced PBLs in almost all cases. The review found that PBLs were focused primarily on activities and outputs, with very poor definition of outcome indicators[53].

2.33 Table 2.1 also shows the instruments that did best within SG. Multiphase operations and global credit operations did well at identifying outputs and their

metrics, as did global multiple works, to a lesser extent. Despite this, the adequate identification of substantive dimensions, as well as outcomes is still low. For example, 83% of global credit operations, which in 2009 were composed mostly of PROCIDADE operations in Brazil, were rated satisfactory or better in outputs, but none of them adequately identified appropriate outcome indicators and baselines. The same is seen in Multiphase operations, where definition of outcomes was still poor.

2.34 The bottom panel of Table 2.1 reports the percentage of projects which were rated at least satisfactory, by country region and responsible Bank unit. The Bank is organized into four country regions. It is also organized in two vice-presidencies responsible for project preparation: VPS and VPP. Within VPS there are four units that are responsible for project preparation, and within VPP there are two. The results show that the Andean and Southern Cone country regions did somewhat better than the other two country regions. The difference between the two groups is driven mostly from the different identification of indicators, where the proportion of satisfactory or better is much larger for the Andean and Southern Cone groups. With respect to substantive dimensions, Andean and Southern Cone did better, but the percentages are very low for all four regional groups, as it is in the identification of baselines.

Figure 2.3 Evaluability over time: Averages by Dimension



2.35 The table also shows that social programs are substantially more evaluable than other programs, as was the case in 2005. For example, 71 and 65 percent of projects prepared by the Social units had a satisfactory identification of outputs and their baselines, compared with 44% and 35% for projects prepared by the Infrastructure units, and 58% and 35% for projects prepared by the Institutional and Finance units (modernization of the state). This difference across units is not as clear in the case of substantive dimensions, given that such a low percentage of projects achieved a satisfactory rating in those dimensions. It should also be noted that with respect to the monitoring framework, projects prepared by the Social division had a much better performance than projects prepared by other Bank units, with 59% of them having a satisfactory or better rating.

2.36 Figure 2.3 tracks evaluability scores by dimension over the three years studied. Although evaluability scores are ordinal in nature, averages were presented in order to summarize results across time; a full distribution of frequencies is provided in the Annex. From 2001 to 2005 evaluability ratings had improved in formal dimensions and had worsened in substantive dimensions. In 2009 the data show deterioration in all but one of the evaluability dimensions. The only dimension in which there is some improvement is in the identification of assumptions and risks.

2.37 The change in evaluability over time coincided with a large increase in the amount and number of project approved. Figure 2.4 details the change over time in both the volume and number of Bank approvals. The data show an increase over time in project approvals in the last three years, following seven years of approvals that fluctuated around 5 billion a year. In 2007 amounts were higher, in 2008 again higher, and once more in 2009, higher. The number of approvals also parallels the trend seen in approvals. The decision to prepare and approve more operations is relevant for two reasons: (i) first, it created additional demands on time and resources—which clearly can have implications for evaluability—and (ii) second, it created a large corpus of assets that would inherit the evaluability deficits documented in this report.

Figure 2.4: Approvals (millions) and projects approved (number)

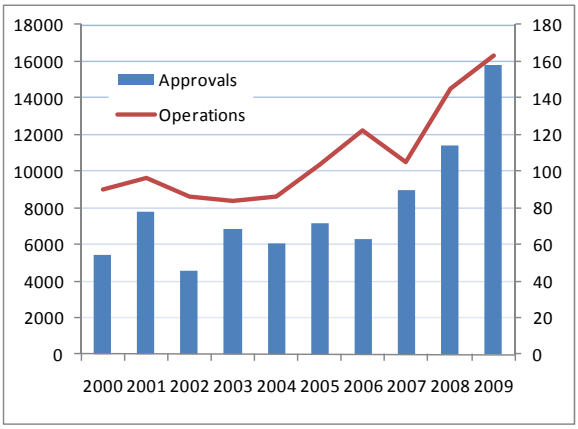
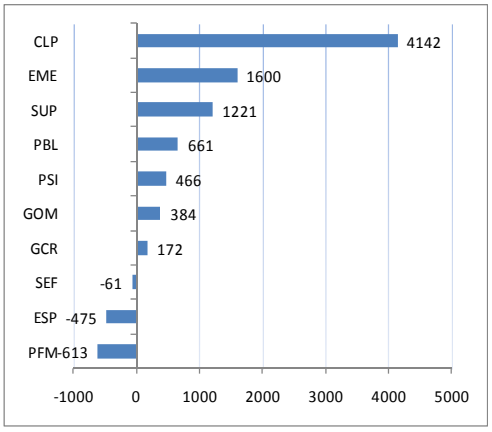
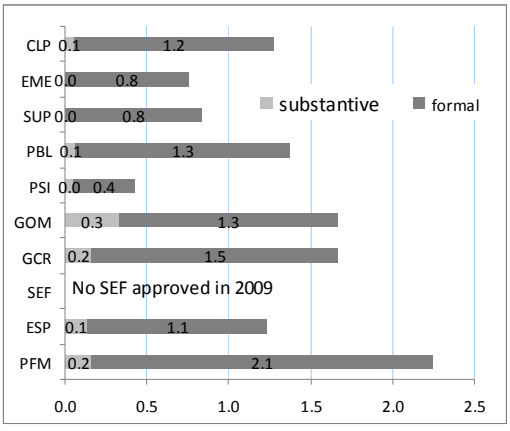


Figure 2.5: Metrics by instrument: Change in approvals 2009-2005 (left) and number of adequate dimensions (right)





2.38 Figure 2.5 shows that the increase in approvals was concentrated in four instruments: CCLIPs (CLP), Emergency Lending (EME), Supplemental Financing (SUP), and to a lesser degree Programmatic PBLs (PLB)[54]. Figure 2.5 also details the average number of evaluability dimensions rated satisfactory or better, by instrument type, and it shows that the instruments that saw an increase in amount were also among the least evaluable.

2.39 The large amount of lending in 2009, coupled with the evaluability deficiencies documented above, creates a portfolio which will be difficult to evaluate in the future. Projects for which the relationship between problems and solutions are not clear, projects with poor diagnostics, and projects inadequate sets of indicators and monitoring frameworks will limit what the Bank can learn regarding development effectiveness from its main asset: its approved portfolio.

Table 2.2: Amounts approved, according to the number of adequate evaluability dimensions

	2001		2005		2009	
N. adequate dimensions	Approvals	Share	Approvals	Share	Approvals	Share
			Substantive Dimensions			
None	*1 696*	*0.19*	*3 893*	*0.57*	*12 189*	*0.88*
One	*2 071*	*0.23*	*1 842*	*0.27*	*2 046*	*0.08*
Two	*2 086*	*0.23*	*0 432*	*0.06*	*0 130*	*0.03*
Three	*2 711*	*0.30*	*0 706*	*0.10*	*0 000*	*0.00*
All	*0 535*	*0.06*	*0 000*	*0.00*	*0 000*	*0.00*
Sum	*9 098*	*1.0*	*6 873*	*1.0*	*14 365*	*1.0*
			Formal Dimensions			
None	*1 197*	*0.13*	*0 533*	*0.08*	*7 103*	*0.46*
One	*1 557*	*0.17*	*0 565*	*0.08*	*1 457*	*0.14*
Two	*2 793*	*0.31*	*1 363*	*0.20*	*2 695*	*0.29*
Thee	*3 296*	*0.36*	*3 042*	*0.44*	*0 805*	*0.07*
Four	*0 108*	*0.01*	*0 552*	*0.08*	*0 677*	*0.01*
All	*0 148*	*0.02*	*0 819*	*0.12*	*1 628*	*0.02*
Sum	*9 098*	*1.0*	*6 873*	*1.0*	*14 365*	*1.0*

*Approvals are in millions of USD

2.40 Table 2.2 attempts to dimension the size of this problem, by presenting the dollar amount of approved resources, according to the number of evaluability dimensions for which ratings are satisfactory or better (a `3' or a `4'). As can be seen, among substantive dimensions of evaluability, in 2005 57% of the approvals were in projects with no adequate dimensions. In 2009 88% of the portfolio approved was in this category, which is equivalent to more than 12 billion dollars. In formal dimensions, the percentage of the portfolio with no adequate evaluability dimensions was 13% in 2001, and 8% in 2005. In 2009 48% of the portfolio approved was in this category, which is equivalent to over 7 billion dollars. The approval of large sums of unevaluable investment generated a contingent liability of projects which are "at risk" for not being able to demonstrate their results in the future.

III. MANAGEMENT'S REVIEW PROCESS

3.1 Project evaluability is shaped by how the Bank reviews each proposal as it moves through the stages of approval up to final submission to the Board. Evaluability is thus one of the outcomes of the Bank's project review and approval process. If there are persisting (worsening) problems of evaluability, the explanation for this must reside in the review and approval process itself. OVE's Evaluability Analysis for 2005 (RE-333), identified problems with the review process, including delegation of responsibility, the dilution of accountability, institutional incentives, and the specific quality control processes.

3.2 Since the 2005 review, Management has made significant changes to the project review and approval process. In 2007, Management introduced both a New Project Cycle (CS-3734), or NPC, and a proposed Realignment of the Bank (GA-232). Both sought to reconcile the need for quality control with the need for an improved timeliness in the Bank's response to country needs. The Realignment document described the perceived problems with existing systems of quality control in the following terms.

The analysis of current practices has identified a pervasive confusion between the roles of quality control and quality enhancement as a key problem in the Bank's present structure and procedures. […] This "cohabitation" of quality control and operational support responsibilities generates a cascade of undesirable consequences: accountability of the operational units for project quality is diluted; the independence of the quality control unit is compromised by their subordination to the pressures of project approval and disbursement; the process of reaching a consensus between operational and safeguards units is very time consuming and focuses on changing document texts rather than on risk management on the field". (GA-230, §5.6 and § 5.7, GA-232).

3.3 To address these concerns, Management proposed a new ethic of individual responsibility for project quality, a new allocation of responsibilities for quality assurance, and a new set of review processes.

3.4 **Individual Responsibility**. The new review process was predicated on an assumption of individual responsibility for project quality. According to the NPC document:

Committees will generally have advisory status, not decision making powers. When projects are prepared, approved, or assessed for their performance, documents will be submitted by individuals who will attest by their signature the direct responsibilities for the quality of the proposals. (§ 2.5).

3.5 In practice, this new ethic has meant that project teams assume greater responsibility for the content of loan proposals, and incorporate external feedback and assessment only to the extent that they deem necessary.

3.6 **Organizational Responsibilities**: The Realignment proposal assigned responsibilities for project quality to a newly created Office of Strategic Planning and Development (SPD) and to the Vice Presidency for Sectors (VPS) according to the following division of labor: *In order to mitigate the development risks of Bank activities, SPD will focus on assuring the quality of the Bank program* (§6.15). […] In contrast, VPS: …*will be responsible for the quality of individual projects and non lending products in order to ensure that products meet standards and adequately identify developmental risks* (§ 6.48). Earlier in the document (§2.5) , VPC is also charged (along with VPS) for quality control of projects. Thus at least three organizational units, VPS, VPC and SPD share responsibility for the quality enhancement function, which does not seem to solve the dilution of responsibility that was diagnosed as significant part of the problem.

3.7 That ambiguities remain in the assignment of responsibility for project quality is demonstrated by the 2010 Development Effectiveness Overview, which reports that SPD (nominally responsible for quality control) in fact also has an extensive role in supporting teams in project preparation. Not only do SPD staff help teams draft their Development Effectiveness Matrices, but they also play a direct operational role in supporting evaluation components of projects. The report notes: "*SPD role may occur at three levels: advising teams on general evaluation questions; reviewing terms of references and/or results matrices and supporting the team in the definition of the evaluation design (sample size, definition of indicators); or a more substantive role, in which a SPD member participates as full team member, responsible for the evaluation component of the project. Currently, SPD staff is involved in over forty projects at different levels, working with eight sector divisions and also with the Multilateral Investment Fund (MIF).* (GN-2545-1, § 2.24)

3.8 It is clear from this text that the problems of "cohabitation" between quality control and operational support functions noted in the Realignment document have reappeared in the Bank's practice following the realignment. This is both a problem of accountability (identified problems are allowed to persist) and a problem for the quality control function as a whole, which may be significantly compromised by the extensive role taken by the control unit (SPD) in the actual production of products whose quality it is supposed to control.

3.9 **Review Process**. The Bank has two formal review processes, one for SG operations and another for NSG. The NPC (CS-3734, September, 2007) defined a new formal review process for SG operations. This process begins with the creation of a Project Profile document (PP), which is subject to internal quality review procedures in the initiating department and then considered by an interdepartmental Eligibility and Risk Review Meeting (ERM). If eligible, the team prepares a Project Operation Document (POD) which is then sent to an interdepartmental Quality and Risk Review meetings (QRR), after which it is sent to the Operations and Policy Committee (OPC) of senior management prior to submission to the Board.

3.10 The 2007 NPC did not apply to NSG projects, which is subject to a review process largely within VPP. Approval procedures for NSG involved: (i) ERM chaired by VPP, (ii) due diligence, which is an internal VPP activity, (iii) a Credit Review Meeting (CRM) in which other departments are invited to participate, (iv) an optional final review by OPC, and (v) Board Approval. Over time, NSG review processes became more interdepartmental, and in particular involved substantial participation from the Bank's Risk Management Group.

3.11 These formal review processes are not informed by clear quality standards for the evaluability of projects. Standards for project evaluability are required so that project teams know what is expected of them and so that the review process can have explicit guidelines as to what is to be reviewed. The lack of such standards in the Bank was pointed out in OVE's first evaluability report (RE-275, §3.7), leading the Board's Policy and Evaluation Committee to direct that specific evaluability standard be "developed promptly" and be incorporated into "guidelines for project design" by June of 2003[55]. OVE's second evaluability report observed that such standards had not been developed by 2005, and the present review reaches the same conclusion.

3.12 The November 2006 Realignment document recognized the need for explicit standards of project quality, and gave the following mandate to the newly created Office of Strategic Planning and Development Effectiveness:

> *the OSPDE will set quality standards and necessary safeguards for both the design and the execution of the Bank's products,[and] monitor and ensure compliance with standards and safeguards (GA-232 § 6.16)*

3.13 While the Realignment document did not specifically mention evaluability as part of project quality, SPD elected to include evaluability as a component within their broader guidelines for assessing the development effectiveness of projects. The "Development Effectiveness Framework" proposed by SPD in August of 2008 incorporated evaluability as one component of the proposed Development Effectiveness Matrix (DEM) to be prepared for each Bank project.

3.14 While recognizing the need to incorporate evaluability into project assessment, the DEM proposed by Management did not establish evaluability standards. Nor did it adopt the evaluability standards used by OVE in its two prior reports. Despite an explicit instruction from the Board that any new instrument "*… should incorporate clear evaluability criteria and clear provisions on accountability, justifying as necessary any departures from the instrument used by OVE*[56] *.."* no justification for not using OVE standards was provided.

3.15 In lieu of standards, the SPD defines the DEM as "*a review checklist of a project's essential evaluability elements.*" The DEM checklist consists of 73 yes/no questions, 42 of which relate to dimensions of project evaluability. The questions start to be answered by project teams at the PP stage and completed after the POD stage. The answers are rolled up into "scores" for 7 sections of the

DEM, four of which (program logic, monitoring and evaluation, economic performance and risk management) are based on the 42 questions related to evaluability.

3.16 Checklists can have value in that they suggest what questions to pose, but they are not a substitute for standards which indicate what would constitute adequate answers to each question. For example, under project logic the DEM checklist asks: "Is the main problem to be solved clearly identified?", while standards would establish what constitutes a "clear" problem statement.

3.17 Of more concern to the review process is that the questions themselves are not presented in the DEM document that accompanies each project through the review process. Instead, scores are created for each section based on the number of yes answers provided. These scores are complete abstractions, and contain no narrative explaining what features of the project are responsible for what scores, and thus provide little material to work with to improve project evaluability at each review stage.

3.18 NSG operations also have something called a DEM, but its content is significantly different from the one applied to sovereign guaranteed operations. The NSG DEM asks no questions regarding three of the four areas in the SG DEM (project logic, evaluation and monitoring, risk management), and focuses almost exclusively on the anticipated performance of the project, both in terms of economic return and possible additionality. In this sense the NSG DEM is not an assessment of the evaluability of the LD, but rather a prediction of what the project will be able to accomplish. The NSG DEM is also not based on yes/no answers, but instead on reviewer judgments on substantive questions such as the impact of the project on knowledge transfer, innovation and improving corporate governance. Because they focus on specific aspects of the project, the NSG DEM encourages a more substantive discussion at each review stage than does the DEM for sovereign guaranteed projects.

3.19 The implementation of the NSG DEM has, however, caused a significant loss of information in NSG project documents. The DEM replaced the existing logical framework and results matrix sections of NSG loan documents, with the result that loan documents themselves no longer define a specific development objective nor provide any logical linkage between project components and desired final outcomes. The DEM narrative does not substitute for such discussions, with the result that it is often difficult to understand precisely what the rationale or logic is behind each NSG project. This issue is in large part responsible for the fall in evaluability of NSG operations.

3.20 OVE has reviewed all of the written materials associated with the formal review processes of both SG and NSG operations. For SG projects, all are subject to internal review at division level, but without formal guidelines, procedures or standards for the review and without producing writing materials. In the absence of formal guidelines, the evaluation found idiosyncratic and informal supervision

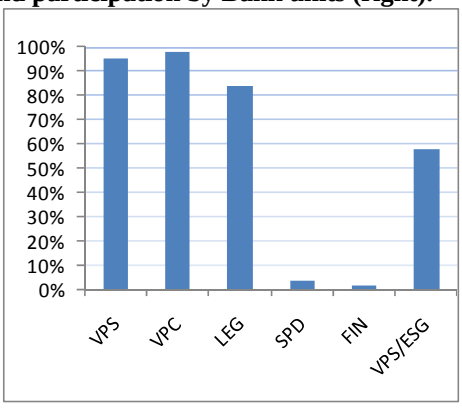
of project quality at the level of the operational vice-presidencies. Different VPS divisions have adopted disparate review practices ranging from external peer reviews (briefly, in the case of social protection), single-person peer reviews, anonymous reviews, supervisory reviews, and, in one instance, no structured review process at all. This occurred in a context of centralization of thematic expertise within the same unit, so that experts from other reporting lines are not available to act as independent reviewers.

3.21 The first level of formal review takes place at the ERM meeting, which is a middle management review chaired by the respective VPC General Manager and includes the VPC Country Representative, VPS divisions, ESG, SPD, LEG and FIN as well as the correspondent country coordinators and economic advisors. The minutes of this meeting are ratified by VPS and by the representative from VPC, jointly.

3.22 OVE's review of all ERM minutes produced in 2009 shows that evaluability issues are generally not considered at this stage. Figure 3.1 shows the frequency with which issues are considered. Apart from risks, in which in 64% of cases there is a discussion, very few considerations related to actual or potential evaluability shortcomings were noted as one of the determinants for the eligibility of projects. Discussions mostly revolved around issues identified in the normative document (PR-1002): the alignment of the PP with country strategies, definition of the financial package, choice of instruments, and fiduciary and environmental risks, as established in the terms of reference of the ERM[57]. Furthermore, participation in this review instance is low and concentrated in VPC (chair), LEG and ESG. VPS does not participate as reviewers. VPS is only represented by the project team.

Figure 3.1
ERM consideration of evaluability issues (left) and participation by Bank units (right).

Issues considered in ERM	% of issues mentioned in ERM
Evaluability	
Diagnosis	19.05%
Objective definition	11.90%
Logic	23.81%
Assumptions and Risks	64.29%
Output Indicators	0.00%
Outcome Indicators	7.14%
Baselines for outputs	2.38%
Baseline for outcomes	2.38%
Monitoring and Evaluation	9.52%



3.23 After eligibility by the ERM, the project team prepares a more detailed Project Operation Document (POD), a process that is also subject to the division level review described above. The POD is then sent for a QRR, which is defined by the NPC as the main formal instance for project quality assurance. The corresponding VPS division distributes the POD to the Country Representative, Country coordinator and economic advisor, LEG, SPD, Procurement and Finance

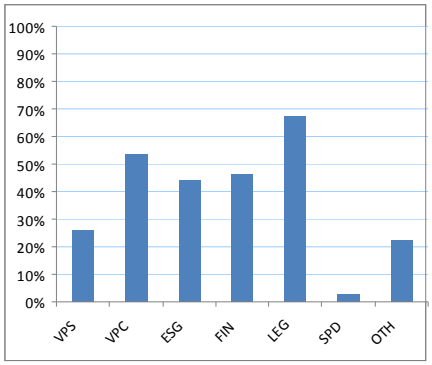
divisions. Unlike ERMs, QRRs are virtual meetings[58]. Project teams prepare a Results and Procedures Report containing the topics raised at the QRR, along with the team's response to comments, which may or may not contain specific actions or changes. These meetings are the responsibility of respective division chiefs in VPS, but in practice project team members are responsible for assembling comments and responding to them.

3.24 OVE's review of all QRRs for projects approved in 2009 shows that evaluability considerations were rarely raised at QRRs.[59] Figure 3.2 shows the frequency distribution of issues considered. Risks, particularly fiduciary and environmental, were considered in 23% of QRRs, followed by issues related to project diagnosis and logic, in 20% and 16% of cases respectively. Rather, topics more frequently discussed included financial and contractual conditions of the loans, procurement and eligibility for disbursements, and other procedures for new instruments. The specific role and corresponding review standards according to their domain of responsibilities for each Bank's organizational unit participating in QRR have not been defined.[60]

Figure 3.2: QRR consideration of evaluability issues (left) and participation by Bank units (right).

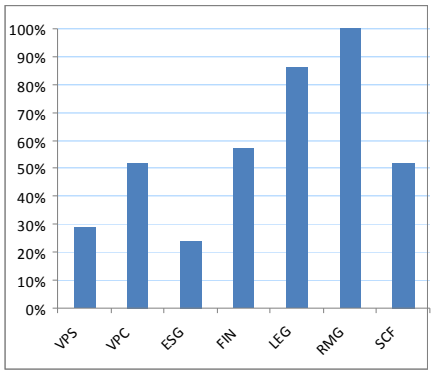
SG QRR

Evaluability issues considered in QRR	% of issues discussed in QRR
Diagnosis	20%
Objective definition	5%
Logic	16%
Assumptions and Risks	23%
Output Indicators	5%
Outcome Indicators	6%
Baselines for outputs	0%
Baseline for outcomes	3%
Monitoring/Evaluation	9%



NSG QRR (CRM)

Evaluability issues considered in QRR	% of issues discussed in QRR
Diagnosis	0%
Objective definition	0%
Logic	5%
Assumptions and Risks	100%
Output Indicators	0%
Outcome Indicators	10%
Baselines for outputs	0%
Baseline for outcomes	0%
Monitoring/Evaluation	0%



3.25 The evaluation also found that participation by Bank divisions invited to the QRRs is low. Participation (comments sent to QRRs) is concentrated in the following departments: LEG (67%), ESG (45%), FIN (46%), and VPC (54%). VPC's participation is mainly represented by country economists/coordinators, procurement officials and operational specialists in country offices; 70% of

comments were at this level, while 30% were at the Country Representative and/or Regional Manager level. Meanwhile, the participation VPS divisions (other than the division hosting the QRR) and by SPD, who presumably would have the most to say in terms of project technical soundness and project evaluability respectively, is low, at 26% for VPS and 3% for SPD. Lastly, the review found that participation is concentrated among rank-and-file Bank staff; participation by division chiefs, or by higher Bank personnel accounts for only 11% of recorded comments.

3.26 The low participation by divisions responsible for quality control and assurance in QRRs has produced an unusual practice, in which Bank divisions which in the past have not taken the role of commenting on evaluability, such as LEG, have increasingly taken on this role (information missing on diagnosis, data inconsistencies, problems related to the logic of the intervention), in addition to commenting on more traditional issues usually in their domain. As a result of these content and participation issues, QRRs have been ineffective as a mechanism of systematically assessing project evaluability[61]. This may suggest a failure of the Bank's matrix structure to provide appropriate technical checks and balances through interdepartmental interaction.[62]

3.27 The formal process of review of NSG operations shared many of the characteristics of the SG review. Participation rates from other departments in the CRM were low. Comments were mainly provided by RMG (the Bank's Office of Risk Management), LEG, and to a lesser degree FIN. NSG review focused (almost exclusive) on credit risk: the discussion is dominated by assessment of project, borrower, sponsor, exchange rate and demand risks. As in the case of SG operations, comments did not focus on evaluability issues. There are no instances of comments related to diagnostics, objectives, expected results, indicators, or baselines, or a monitoring framework for results. Bank non-financial aditionality, or market failures are also not discussed. In terms of financial additionality, 15% of minutes show a discussion of this issue. Finally, SPD was generally not involved in the review process of NSG operations.

3.28 An important point of difference between the two review processes is that NSG reviews are often decision-making events. QRR meetings for SG operations virtually never make substantive decisions regarding the content of projects. In contrast, during CRMs important decisions are made. These decisions are taken based on risk and other related findings regarding the project structure, acceptable levels of Bank exposure, and requirements for collateral or other issues regarding assets. CRM reviews thus have real consequences for projects, unlike QRR meetings on SG operations. However, there is no evidence that evaluability concerns have been the basis for decisions taken at the CRMs.

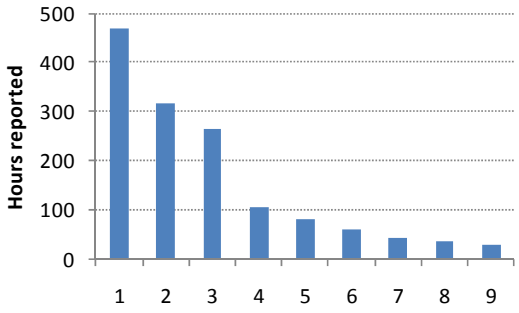
3.29 In addition to the formal review processes described above, there is another process that is intended to help project teams to deal with evaluability issues. As noted in the Development Effectiveness Overview:

In 2009, the DEM became an integral part of the project cycle, from design to approval: each intervention was rated for evaluability at entry prior to Board approval. The DEM is completed by the project team and reviewed by SPD, which may provide support to teams to improve the project evaluability at entry. (GN2545-1 para 2.7)

3.30 According to interviews with project teams, DEMs are frequently discussed informally with SPD, and the DEM that is sent to the QRR contains the scores as they have been "validated" by SPD. Because of the checklist approach used in the DEM, the validation exercise involves a limited review of whether individual questions should keep the "yes" or "no" assigned by the project team or be changed to the polar opposite value. It is possible that conversations between SPD and the project teams go beyond the simple validation of checklist answers, but such information is not recorded. The formal record of QRR minutes, however, reveal no discussion at those meetings regarding DEM related issues, and according to data provided by SPD, only 16% of the projects validated by SPD were later revised by their respective project teams[63].

3.31 Some insight on the nature of the interaction between SPD and project teams can be gained from an examination of time spent on the process. OVE examined the TRS data on hours spent on DEM reviews by SPD staff in 2009. Dividing these hours by the number of project DEMs for 2009 and by the number of retrofitted DEMs in 2008, results in an average of 8 hours per project. This is, however, an overstatement of time dedicated per project DEM, as the TRS categories include other activities unrelated to the review of project DEMs[64]. This work was highly concentrated in just a few reviewers. Three reviewers, account for 75% of time expenditures in the exercise. The remaining six reviewers accounted for 25% of expenditures (see figure 3.3). Note that this time includes only time explicitly reported to evaluability products and not time reported under the TRS codes of the specific projects. Note that if hours were reported under other codes, such as the TRS code for the specific project, it would not be reflected in these numbers[65].

Figure 3.3:Allocation of time in DEM-related activities for SPD Staff



3.32 **<u>OPC Review.</u>** The last management review stage is the Operations Policy Committee (OPC), which can review both SG and NSG operations. The NPC principle of assigning clear and unambiguous responsibility (and accountability) to individuals was not applied at the level of the OPC. In fact, top management responsibility is not assigned to the EVP but to the Committee (OPC) which according to the NPC holds an advisory role. Loan Documents are submitted to OPC for consideration by non-objection. The Committee only discusses an operation

where VPC or VPS requests guidance on a critical issue, or where a policy related matter requires senior management consideration (§3.11).

3.33 In 2009 almost all projects were presented through simplified procedure. Of those that did not, an OPC was convened in very rare instances. In all of 2009 only 10 projects approved during the year (or 7%) were presented for the discussion at the OPC.[66] OPC meetings were not called for to address issues regarding the evaluability of projects or development effectiveness issues. In the 10 cases where a meeting took place there was in fact little discussion of issues related to evaluability.[67]. Indeed, the review of OPC minutes shows that in most cases the discussion focused on procedural issues such as recognition of past expenditures, disbursement commitments negotiated with government previous to the OPC consideration or the nature of the lending instrument. And in the 10 cases where OPC held a meeting, projects were cleared for submission to the Board with few changes.

3.34 The Bank's Development Effectiveness Framework prescribes that: "*..if an operation is scheduled for discussion at the Operataions Policy Committee, SPD will provide its technical opinion on the evaluability of the operation at that time.*" (GN-2489, Appendix 1, paragraph 2.16).

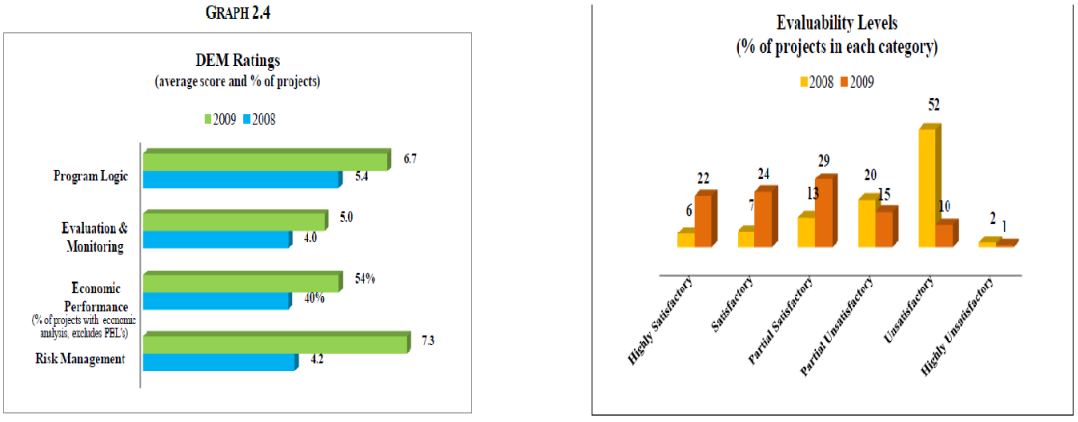
3.35 Of the 10 projects sent to the OPC during 2009, the minutes show no discussion of the DEM, and make no mention of whether SPD rendered its required technical opinion on the evaluability of the project.

IV. REVIEW OF THE DEVELOPMENT EFFECTIVENESS MATRIX

4.1 OVE's 2010 work program included an oversight study on the implementation of the new Development Effectiveness Matrix approach. Because the DEM has become management's preferred approach to assessing project evaluability, this chapter will review the DEM results for the 2009 portfolio, compare those results with those obtained by OVE, and explore issues of method related to each of the two different DEM instruments (SG and NSG).

4.2 The previous chapter demonstrated that the DEM has not yet been effectively integrated into the Bank's processes for approving loans, that evaluability concerns are rarely discussed as an important aspect of project design at any stage in the approval process, and that SPD staff do not participate in the formal reviews of projects at the QRR. Yet despite the absence of a close association between the DEM and the approval process, SPD uses the results of its DEM validation exercise to produce a summary report which purports to measure the overall evaluability each year's cohort of approved projects. For project years 2008 and 2009, this summary is provided as part of the Development Effectiveness Overview (GN-2545-1). For each year, the document computes "evaluability ratings" based on the yes/no checklist answers in four sections of the

Figure 4.1: Average DEM Scores and DEM Ratings, 2008 vs. 2009



Source: DEO, 2009.

4.3 The left panel of Figure 4.1, (taken from GN-2545-1) shows DEM scores for these four sections, while the right panel shows a distribution of projects based on a single composite 'evaluability score' which is constructed by SPD from the scores of each of the four dimensions. The underlying data for both charts are the yes/no answers to specific questions which are weighted to sum to a score of 10 for each dimension shown in the left panel, and then re-weighted into six categories to produce the single composite score shown in the right panel. (For more on this method, see paragraphs 4.15 to 4.17 below).

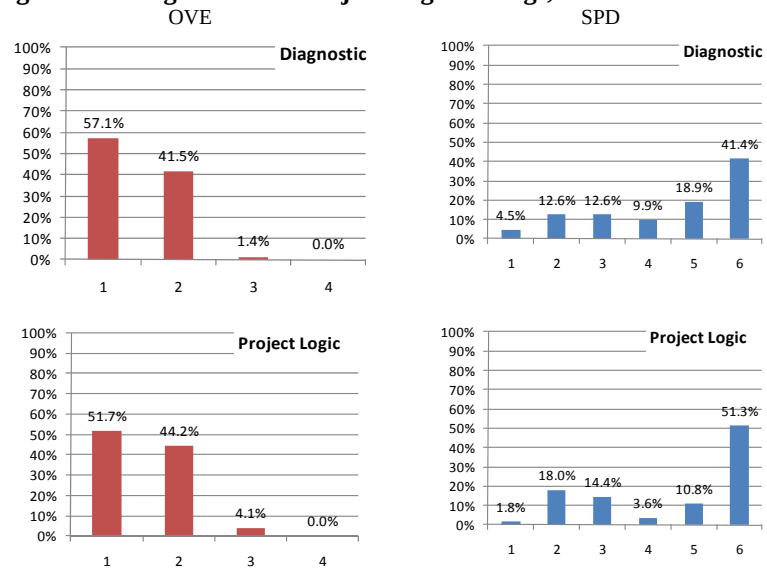
4.4 These charts suggest that project evaluability has improved significantly between 2008 and 2009 in each of the four DEM sections, and that 75% of projects approved in 2009 had DEM scores of 5 (partially satisfactory) or above. In subsequent discussions with the Board, a DEM score of 5 came to be defined as the minimal threshold for adequate project evaluability, a threshold that was subsequently incorporated into the document supporting the ninth capital increase. By this standard, only 25 percent of the Bank's projects have unacceptable levels of evaluability.

4.5 These findings are strongly at odds with the results of OVE's evaluability assessments which were presented in earlier chapters of this report. To help clarify the reasons for these differences, OVE obtained from SPD the details of their ratings for 111 sovereign guaranteed projects approved in 2009 and compared those with OVE evaluability ratings for the same projects.

4.6 As explained in Chapter 1, OVE does not assign a summary evaluability number to each project, and instead reports evaluability scores on each of the 9 dimensions rated by OVE. Thus each project has 9 individual evaluability scores as opposed to one from SPD. This is a deliberate methodological choice, since the compression of multiple scores into a single number involves considerable

loss of information, and assigning weights to the individual dimensions could lead to a non transparent exercise in judgment that can have huge implications for the final scores.

Figure 4.2: Diagnostic and Project Logic Ratings, OVE and SPD



4.7 Both the OVE method and the DEM, however, have similar questions related to diagnosis and project logic. Although the DEM does not provide multinomial ratings for each dimension, utilizing the same method and weights that SPD uses to classify the total DEM scores into six categories, it is possible to generate six categories of DEM scores only for the questions related to diagnostic and logic. SPD scores on a 6 point scale, while OVE uses 4 points, but since they both use even numbered rating scales, each has half its categories below satisfactory levels and half above. The charts above compare evaluability scores on the dimensions of diagnosis and project logic.

4.8 It is obvious from the distribution of scores that the two methods produce very different results and give very different perceptions of the evaluability of Bank projects. These findings pose an important challenge to the Bank. As part of the IDB-9 capital increase negotiations, Governors issued clear and detailed instructions which place evaluability assessments at the heart of the project review process:

Governors endorse a further strengthening of the Operations Policy Committee (OPC) and the programming process, by the President of the Bank and Senior Management, to ensure that projects meet minimum evaluability thresholds. In this respect, Management will amend operational procedures, by end of Q3 of 2010, according to the following criteria: (i) all SG and NSG projects must be rated for evaluability; (ii) the evaluability score includes only the dimensions of evaluability of the DEM; (iii) SPD will support teams in meeting evaluability standards from project profile to project proposal, and will validate the final evaluability score for OPC consideration: RES will review the existing

methodologies for scoring evaluability to determine any required improvements; OVE will report annually to the Board of Executive Directors on project evaluability (ex-ante), as well as validate achieved results in completed projects (ex-post): (iv) a minimum evaluability threshold of 5 will be required for all operations to be submitted to the Board of Executive Directors

4.9 Given that OVE is tasked with annual reporting on project evaluability, and RES is tasked with improving the DEM methodology for evaluability assessment, it is important that the two assessment processes work to converge on matters of method in establishing evaluability standards for the Bank. As a contribution to this process, the remainder of this section will make a number of observations on the differences between OVE's evaluability assessment and that done in the DEM.

4.10 OVE approaches evaluability assessment as a fundamentally qualitative process, requiring detailed exploration of the nuances of each specific project being reviewed. Qualitative assessment begins with a project narrative of some 4-6 pages that outlines potential evaluability issues linked to data provided in the loan document. These narratives are subject to peer review by a team within OVE, which encourages convergence on common approaches to rating each evaluability dimension. Only after this qualitative peer review are dimension scores assigned.

4.11 OVE's fundamental concern in the project narrative is an answer to two questions: "does the project define clearly the problem to be addressed," and "Is the proposed solution logically connected to the problem?"

4.12 The DEM approach is quantitative, based on yes or no answers to a checklist of questions. No project narrative is prepared that addresses evaluability concerns, and no internal peer reviewing is done within SPD. In OVE's opinion, a simple checklist is not well matched to the task of assessing ex-ante project evaluability. Checklists are used extensively in airlines and hospitals to ensure that pilots and surgeons comply with all of the steps in a well-established routine that is extensively researched and documented[68]. Checklists thus help reduce the error rate in a long series of related tasks where complacency and boredom are the major threats to quality.

4.13 In the case of Bank projects, however, each situation is unique, and project teams are engaged in a process of creative discovery where checklists are much less helpful. For project design, the threat to quality comes from an incomplete understanding of the situation, not from a failure to perform an established routine.

4.14 To take a concrete example, in the section related to "program diagnosis," the DEM checklist asks: "Is empirical evidence of the problem provided?" The question is relevant, but is only a starting point. Such a question could be answered yes if a single piece of data related to some minor subcomponent was available, while a no would signify the complete and total absence of any data on

the problem. This kind of forced choice situation has two undesirable consequences: it creates a bias toward yes responses, and it discourages deeper inquiry into the issue at hand, since a yes/no answer requires minimal information[69].

4.15 After each question is answered yes or no, the DEM method computes evaluability scores mechanically for each of the four dimensions shown in Figure 4.1 above. The method requires that each area have a maximum score of 10, but since there are different numbers of yes/no questions in each area, questions in some areas carry larger point values than others. These implicit weights are not conferred based on the significance of the question, but only on the number of other questions in each area. In addition, a second weighting exercise is required to produce the single composite evaluability score shown in the right panel of figure 4.1. This exercise gives two of the areas (program logic and evaluation) weights that are twice as large as those given to economic performance and risk management in the determination of the final composite score.

4.16 These two adjustments produce an extreme dispersion of weights for each question, with the heaviest question having a weight that accounts for 42 times the weight of the lightest. (The individual questions and their final weights are presented in the Annex) Such weighting imbalances are not based on analytic work, and can produce perverse effects in the scoring of individual projects. For example, because of heavy weightings assigned to some variables such as evaluation design and risk rating, a project could achieve a minimum composite score of 5 while completely lacking a diagnosis of the problem, a logical intervention to solve the problem, or any sort of results matrix.

4.17 The issue of question weights becomes especially important in the context of establishing an overall "minimum evaluability threshold" as recently endorsed by the Governors. Current practice in the DEM is to produce weighted average scores and to set 5 as the minimum. The dispersion of question weights means, however, that projects with very different characteristics can achieve average scores of 5, even if they have very little in common in terms of the different evaluability dimensions. An alternative approach would be to assign minimum thresholds to each review category and to define an "acceptable" project as one having at least a minimum score on each dimension.

4.18 Yes/no questions force black and white answers on situations that are frequently shades of grey. For example, the first question under diagnostics asks: "*Is the main problem to be solved clearly identified?* From a reviewer point of view, the clear incentive in such a question is to answer "Yes", and in fact, 93% of projects were so rated in their respective DEMs. The more complex and multi-dimensioned the question, the greater the incentive to provide a yes answer.

4.19 In addition, the incentives created by the yes/no format may create inconsistencies between the judgments made by reviewers and the information actually contained in the project documents. To check on this possibility, OVE conducted a validation review of 33 projects to test for consistency between the DEM score and data in the loan documents. This review found that in 41.7% of the cases the evidence available in project documents is inconsistent with the ratings assigned by SPD. The percentage of instances in which the loan evidence is not consistent with the DEM ratings ranges 15.2% to 73.2% depending on the areas. These percentages are particularly high in the areas of "diagnosis" and "results matrix quality". In these areas, the DEM provides ratings that, in most of the cases, are not compatible with the evidence contained in the loan documentation.

Table 4.1: Results of the SG DEM validation

Criterion	Percentage of instances in which evidence is contrary to DEM rating
Section 3. Program Logic – Area Rating	
Program Diagnosis	73.2%
Proposed Solutions (@ PP)	26.3%
Proposed Solutions (@ POD)	59.1%
Results Matrix Quality	54.5%
Section 4. Evaluation & Monitoring – Area Rating	
I. Evaluation	25.1%
II. Monitoring	22.2%
Section 5. Economic Performance –Area Rating	
Economic Rate of Return (ERR)	16.7%
Cost-effectiveness	15.2%
Section 6. Risk Management – Area	
Risk Matrix Score	72.7%
Mitigation Matrix Score	42.4%

4.20 Finally, the DEM assessment questions miss important qualitative dimensions of project evaluability that are not easily captured in yes/no questions. In the diagnostic section, for example, questions are asked regarding the existence of a problem statement and the existence of data, but do not ask whether the diagnosis makes sense. It assesses the presence of a diagnosis, but not its quality. Similarly, there is no required assessment of the consistency between the problem diagnosis and the solution proposed. If it has a problem statement, it gets a check in that section. If it has an intervention description, it gets a check in that section, whether or not there is a meaningful connection between diagnosis and action.

V. MANAGEMENT'S RESPONSE

5.1 As part of preparation for this report OVE distributed the draft text of the report, as well as all the *Notes* for projects reviewed in the evaluability exercise, for peer review by Management. Management's comments were important to improve the quality of the final report. The comments also provided additional insight into the possible causes of evaluability problems documented.

5.2 OVE received responses for 111 of the 147 *Notes*. In reviewing these responses, OVE concluded that changes were warranted in the text of 33 of the 111 *Notes*.. However, the changes were mostly minor, and OVE's review concluded that modifications were warranted in the ratings of 5 projects. The *Notes* annexed in this document reflect the changes made. The remaining comments received fell into one of two broad categories: (i) comments regarding policy or resource constraints and their implications for evaluability; and (ii) comments regarding difficulties in meeting evaluability requirements. Within these two broad categories there are five distinct types of comments received.

Policy or resource constraints and evaluability

5.3 Table 5.1 presents the frequency distribution of each of these comments. Seventy-one percent of project teams argue that limitations of the loan proposal format—including the number of pages—or other limitations associated with the NPC, such as time and budget constraints, were issues in meeting evaluability standards. This also includes NPC rules allowing operations to define some of their components in the future. In 62% of cases project teams argued that projects did not meet a specific evaluability standard either because these did not apply to them, or because they argued that they were too high. The argument of applicability of standards was made most often in the case of emergency loans, as well as private sector loans and transportation loans.

Table 5.1: Frequency distribution of comments received (percentage)

Difficulty meeting standards			Policy or Resource constraints			
Overall	Standards not applicable to the type of project	Standards are too high	Overall	Information will be collected or clarified during execution (Evaluability deferred)	Information is not presented or presented in Annexes due to space limitations	Information is not presented due to resource limitations
62%	46%	41%	70%	34%	56%	16%

5.4 Of the comments related to limitations imposed by policy or constraints (NPC constraints), the most common one was that of page limitations, which in turn implied that additional information required in order to establish the evaluability of the project was in the project annexes. This is seen in 56% of cases. However, such comments were not accompanied by indication as to where specifically in the annexes the information was to be found. Since annex materials are generally not referenced or footnoted in the POD, it is not surprising that such specific references are missing from Management's comments on OVE's notes. The absence of specific citations make the very lengthy annexes neither an effective support to the review of the POD by the Board, nor the review of the evaluability note by Management. OVE has read every annex for every project, and finds wide variance in the quality of the data, ranging from high quality analytical work which is directly related to the project, to irrelevant extraneous material with little analytical content. It is unclear to what degree this is due to failings in the normative regarding Annexes, and to what degree this is due to the lack of

oversight of quality—a problem already documented in chapters III and IV of this report.

5.5 The disorganized presentation of information and evidence is not a minor issue; it has implications for the understanding of the operations approved, and thus for their evaluability. By imposing high information search costs, this issue can also be an impediment for the optimal decision-making process of both Bank Management and the Board regarding project design and approval. The high search costs have an additional consequence, in that they make the oversight and monitoring of Bank programming difficult, a constraint which can limit both accountability and transparency.

5.6 In addition to constraints of space, 16% of responses mentioned constraints of time and resources as an explanation for observed evaluability problems. The NPC introduced specific restrictions regarding project preparation. These restrictions were implemented across sectors, with little room for flexibility between projects that require fewer or more resources. Assessing the extent to which NPC time constraints—or resource constraints more broadly—are responsible for evaluability issues is beyond the scope of this evaluation. However, the fact that this constraint was raised in a significant portion of projects would call for additional review of the issue by Management.

5.7 The NPC explicitly allows for refinements in design during execution, and in 34% of the comments, project team members explicitly mentioned that the project content would be completed at a later stage, with some comments referencing the NPC normative. While the current NPC normative recognizes that design work may be incomplete at the time of project approval, it also states that in such cases: "*the document will specify what is missing for the project to achieve maturity and exactly how those gaps will be filled.*" (CS-3734) For projects where Management's comments indicate that evaluability will be completed at a later date, the PODs do not contain the specifics regarding gaps and proposed remedies that are required by the NPC normative.

5.8 As with the case of Annexes, this is not a minor issue. Management and the Board are required to make decisions based on the information presented to them. If important design elements have been deferred, and are not made explicit, and if mechanisms with which these gaps will be filled identified, decisions may be taken under a set of incorrect assumptions. Again, this goes to issues of both accountability and transparency in the institution.

Minimum requirements and evaluability

5.9 One of the most common comments received was regarding the standards by which the evaluability exercise evaluated projects. In 46% of cases comments indicated that standards did not apply, and in 41% of cases the comment indicated that the standard was too high.

5.10 The issue of applicability of standards is a policy issue that cannot be resolved by OVE. Both the Executive Directors and the Governors have indicated that project must meet evaluability standards. If Management desires to exempt some types of projects from these standards, it should propose such changes to the Board.

5.11 The issue of excessively high standards related primarily to diagnostics, where project teams felt that both problem statements and proposed interventions did not need to be justified by specific analytical work in context since they were adequately covered in existing general literature. OVE's reviews will accept such generic references, but only in cases where the project design and results matrix contains indicators specific to the project context, in order to be able to determine ex-post whether the generalized response predicted in the literature was realized in this specific context.

VI. CONCLUSIONS AND RECOMMENDATIONS

6.1 This report has demonstrated a curious paradox: the term "evaluability" has become central to the Bank's understanding of how projects should be designed, but the actual evaluability of projects has deteriorated markedly when measured using a consistent methodology applied to three cohorts of projects: those of 2001, 2005 and 2009.

6.2 The report found that the Bank's oversight system is not considering evaluability in its review processes. This extends from the QRRs to the OPC. Minimun evaluability standards were not defined in the review process. Assessment of QRR and OPC minutes shows that the issue is rarely brought up. Meanwhile, the Bank implemented a different methodology for evaluability assessment in the DEM, without justifying its departure from OVE practice. DEM assessments provide much less information on the specific characteristics of projects and have a strong bias toward reporting positive findings. The review also found that DEM scoring was not based on evidence and was systematically different from the results found by OVE when replicating the DEM exercise. Furthermore, preparation of DEM assessments has involved relatively little staff time, and the assessments themselves have not provoked discussion of evaluability issues at any stage in the review process. In sum, the DEM evaluability assessments were not well connected to the processes that actually generate the Bank's portfolio, and the ordinary Bank review processes did not take into consideration evaluability issues.

6.3 In the IDB-9 agreement, Governors instructed the Bank to improve the evaluability of Bank projects. As part of that process, OVE would make the following recommendations:

- Recommendation 1: The review of "existing evaluability methodologies" by RES should include the following elements: (i) a rigorous and systematic review of best practices specific for ex-ante evaluability and quality control of complex

interventions, such as development projects. This should include a theoretical and empirical review of the literature and practice, including issues regarding methods and measurement, implementation processes and mechanisms, and organizational design and prevailing organizational incentives; (ii) a pilot application of alternative evaluability methods to the IDB context, so that real-world performance of competing instruments can be observed; (iii) a validation of the accuracy and quality of the current implementation of the DEM vis-à-vis evaluability questions and minimum standards. In carrying out this review, RES should observe the 2005 Board directive to compare Management's method with OVE's, "*justifying as necessary any departures from the instrument used by OVE"*[70]. Following the Governor's recommendation, OVE will continue to report annually on the evaluability of Bank projects.

- Recommendation 2: Once an adequate evaluability method has been adopted by management, the Bank's internal review process should be organized so as to ensure that the method is applied comprehensively to all Bank interventions. This will require that: 1) sufficient staff time and resources be applied to evaluability reviews; 2) evaluability assessments contain specific narrative on each individual project, and that these narratives form the basis for a thorough peer review of the evaluability of projects. NSG operations should be reviewed for evaluability using the same criteria and review process.

- Recommendation 3. Evaluability is a judgment regarding the clarity and precision with which a project identifies Bank intent. Intention to solve some development problem must be a component of all loan proposals, including NSG operations. The Bank needs to restore the practice of having development intent clearly expressed as a part of every NSG operation. At a minimum, NSG statements of intent need to incorporate the instructions from the Governors to: "*identify and address market failures that justify support from the Bank*" (paragraph 3.33). This had been a component of NSG results frameworks until the 2007 implementation of the NSG DEM, and such frameworks should be brought back into the revised DEM, to be developed by RES.

- Recommendation 4. Evaluability assessment cannot remain a shadow process, disconnected from the formal review of projects. The formal review process for sending loans to the Board needs to incorporate evaluability at each stage of the review. Any formal review, including virtual reviews such as QRRs, needs to have a section dealing with evaluability issues, together with an assessment that the project meets minimum evaluability standards. Waivers of these standards, as contemplated by the Governors, should be explicitly proposed by senior management.

- Recommendation 5: The Realignment document described clearly the problems associated with "cohabitation" between quality control and operational support functions, yet a clear distinction has not been maintained between these two functions in recent Bank practice. This may be an important contributor to the

problems found in the project review process noted in this report. OVE recommends that the division of responsibilities indicated in the Realignment document should be observed, with VPS taking responsibility for supporting project teams and developing high quality projects, and SPD having only responsibility for quality review and control.

- Recommendation 6: The evaluability assessment and the response by Management identified a number of issues, including (i) problems in clarity of operations due to the format in which information is organized and presented to the Board, (ii) a perception among project teams that resource and time constraints are often a limitation to project evaluability; and (iii) a lack of common agreement in the institution regarding what projects should document, measure and demonstrate, and how to accomplish these tasks. To address these issues OVE recommends: (i) that Management develop and implement a clear standard for content and reporting for operation Annexes, and a standard for referencing this content in the POD; (ii) that Management revise the procedures established for loan preparation to allow teams greater flexibility in time, page length, and resources if required to prepare fully evaluable projects; and finally, (iii) that Management review its training and project preparation policies so that there is clear guidance regarding how projects should be designed in order to demonstrate their results, including how they should define their expected outcomes, and how they should assess if these benefits have materialized.

ENDNOTES

[1] For a discussion on evaluability standards that existed at the time see Kostoff, R.N. "*The Handbook of Research Impact Assessment*", Seventh Edition, Summer 1997, Office of Naval Research. For examples of implementation of this method in other contexts see Wholey, Joseph S. *Evaluability Assessment: Improving Evaluation, Management and Performance*. General Accounting Office, 2002.

[2] For a discussion of the issues associated with peer review mechanisms see N. Rons *et. al.*, "Research evaluation per discipline: a peer-review method and its outcomes", *Research Evaluation*, Vol. 17, No. 1, 2008, pp. 45-57. See also the recommendations of the Boden Report on Peer Reviews in the UK's Research Councils, and Kostoff, 1997, *op.cit.*, pp. 49 ff. For a discussion of problems associated with peer review see Foltz, Franz. A. "The Ups and Downs of Peer Review: Making Funding Choices for Science", *Bulletin of Science, Technology and Society*, Vol. 20, 2000, pp.427-440. For a discussion of conditions that should be met in constructing peer review processes see Ormala, Erkki, "Nordic experiences of the evaluation of technical research and development", *Research Policy*, Vol. 18(6):pp.333-342.

[3] These review processes rely on a two-step process, in which a panel of independent experts produce peer reviews of proposals, rate proposals on a multinomial scale regarding the relevant criteria, followed by a second step in which these reviews and ratings are consolidated. This is the method used, for example, by the National Institutes of Health (NIH). Similar processes are also applied by almost all major funding agencies, including the National Science Foundation in the US, the Research Councils in the UK, as well as institutions in the LAC region, such as CNPq in Brazil, Argentina's CONICET, Mexico's CONACYT and Chile's FONDECYT. The World Bank applies a similar method in its review of the quality of projects at entry (QUAG group).

[4] See Kostoff, R.N. "Research Program Peer Reviews: Purposes, Principles, Practices, Protocols. Office of Naval Research, Arlington.

[5] The number of response categories is a topic which has generated its own line of research in fields of education research, psychology, medicine, and marketing. There has been no research of this issue in the context of evaluability, but for a discussion in psychology see Preston, Carolyn C. "Optimal Number of Response Categories in Rating Scales: reliability, validity, discriminating power, and respondent preferences", *Acta Psychologica*, 104, 2000, pp.1-15. For applications in business, see Madhubalan Viswanathan et. al., "Maximum versus meaningful discrimination in scale response: Implications for validity of measurement of consumer perceptions about products", *Journal of Business Research*, vol. 57, no. 2, 2004, pp. 108-124. In general, the analysis of the optimal number of ratings recognizes the existence of some form of tradeoff between amount of retrievable information and the reliability of it. Although there is no consensus as for the optimal number of categories, most of these studies find a loss of validity when using fewer than 3 categories and typically find negative effects on reliability when categories exceed 7.

[6] Documents used for review and ratings include: Loan Document (LD), loan annexes, and in the case of Multi-Phase loans, Programmatic PBLs or CCLIPs, the original LD. References consulted that were not included in annexes or in the LD were only assessed to the degree that they were reflected in the LD. For example, the risk review matrix, which is not incorporated as a standard annex, was only considered to the extent that its findings were incorporated in the LD.

[7] For example, in the Jamaica Roads program (JA-L1027) the document does not identify the extent of disrepair of the roads network, does not identify what the funding requirements would be to maintain the roads network, and does not describe or provide evidence of the institutional challenges that the sector faces, so it is not possible to dimension the magnitude of challenges faced.

[8] For example, in the case of Belize competitiveness program, the loan finances the public provision of agricultural services, yet the diagnostic does not identify these services as a bottleneck for productivity and growth. In fact, the challenges that the LD does identify (e.g. poor policy environment) are not prima facie consistent with what it is financing (subsidized extension services to farmers). In the case of a Bolivian loan to promote child and maternal health (BO-L1036), the loan is proposing a regional and ethnically inclusive approach to maternal health, but the diagnostic does not provide a regional or ethnically-specific assessments of the causes of health outcomes, such as maternal and child mortality. In the case of a housing program for the "majority" in El Salvador (ES-L1035), the document identifies the lack of a mechanism for publicizing program requirements as the cause of the lack of housing credit, but provides

little evidence to suggest that it is so. The LD does not assess the role that alternative (and potentially more plausible) explanations for lack of credit, such as the lack of land titles or the possible presence of too many burdensome bureaucratic steps.

[9] A good example is the social policy PBL in Belize (BL-L1004). Although the document presents an overall description of the sectors and their problems—mostly in 600 pages of annexes—there is actually no data on some of the central issues, such as why certain groups receive services and other do not, and the effectiveness of services for those who receive it. The result is that the program makes statements regarding central development challenges without the necessary evidence.

[10] This is the case in other NSG loans also. For example, the operation providing credit for Banco Agricola in El Salvador does not discuss the market failures associated with the lack of micro- and medium-credit in El Salvador. An analogous problem is seen in the case of BBVA in Panama (PN-L1049): there is no diagnostic of the nature of the credit constraints facing SMEs.

[11] This is the case, for example in Mexico's training program (ME-L1084), where possible beneficiaries are defined as the "the unemployed and those possibly facing layoffs or pay cuts." But the LD does not mention if all the unemployed are eligible, nor does it identify how one would go about identifying workers at risk of losing their jobs. In other cases projects use concepts such as "vulnerable population", "indigenous", "road users", "poor neighbourhoods", etc, without a clear identification of who they are, or their defining characteristics. See, for example, BL-L1004, BO-L1034, BR-L1051, NI-L1055. This is also seen in the case of NSG operations. In the case of the loan to an Ecuador mortgage lender (EC-L1031), the diagnostic mentions "low-to-middle" income as potential beneficiaries, but provides no specificity as to what this means in terms of observable characteristics.

[12] In the case of ES-L1035 the beneficiary population is not clearly defined, so that according to the LD data a full 82% of the population would qualify. In some cases, the LD does not clearly identify benefits beyond the private benefits that would accrue to the borrower. This is the case, for example, of NI-L1045 and CH-L1056.

[13] This is seen, for example, in the loan supporting Mexico's *Oportunidades* CCT (ME-L1067). Despite the program's experience with data generation and evaluation in conditional cash transfers, and the Bank's involvement in this experience, the diagnostic does not utilize these resources to validate the intervention model, particularly in the case of the transmission of inter-generational poverty, where there is no evidence provided of how this has changed between program cohorts and their parents.

[14] In the case of a rural roads program in Brazil (BR-L1231), the LD does not provide the required information in order to identify what did and did not work and why, despite the fact that an impact evaluation was done in 2008, according to the LD. In the case of NSG operations this is also seen. For example, in the case of a food network program, the loan will finance the expansion of a specific food retailer's development model (Mi Tienda model), yet the LD does not describe the findings of a previous pilot experience in order to inform the expansion plan.

[15] In the case of an integrated public transport system in Bogota (CO-L1076), the LD identifies as its objective "*to support Bogota, in the implementation of its Integrated Public Transport System (IPTS) through technical and institutional assistance to coherently coordinate and ensure the viability of the IPTS projects with their various subcomponents*". The focus is on the implementation of a transport system and not on the results that this system will produce.

[16] In the case of a financial sector reform loan in Guyana (GY-L1021), the objective is stated as "*contribuir (a corto y mediano plazo) al fortalecimiento del sector financiero y a la mejora del acceso de las empresas y los particulares a los servicios financieros. En modo especial, el programa ayudará a consolidar la capacidad de supervisión del Banco de Guyana, mejorar el acceso a los servicios financieros, aumentar la transparencia en el sector financiero dando a conocer al público datos financieros y aumentar la eficiencia del sistema de pagos. Asimismo, el programa ayudará a fortalecer la capacidad de las autoridades para combatir el lavado de dinero*." The usage of "contribute to the strengthening" and "help in the consolidation" muddles the clarity of the objective.

[17] In the case of the green bank proposal in Panama (PN-L1056), the LD contains 14 different statements of intent that range from company financial results to the "*transformation of the Panamanian economy into a low carbon development path.*" This set of objectives makes it difficult to determine what the intervention hopes to realistically accomplish.

[18] For example, in the case of an ethanol operation in Peru (PE-L1082) the project's viability depends on the ability to export ethanol to other markets, such as Europe and the United Status, but the LD does not discuss necessary conditions for this to take place, such as the customs and regulatory framework required. This is particularly relevant given that the project does not have a purchasing agreement in place. Another illustrative NSG example is the case of ME-L1066 Puerta de Hierro. Here the justification for the Bank to finance a private-owned hospital is the expansion of the supply of medical services, especially for "outsourced" publicly insured patients, including low-income families affiliated to the Seguro Popular. The Company expects to increase the number of publicly insured patients in the two hospitals combined from 0 to 22% by 2013. However, the LD does not provide evidence that the regulatory preconditions for increased access for "outsourced" patients exists. In fact, references to these preconditions are in the future tense, suggesting they have not yet materialized. (e.g. par. 4.9 refers to *"federal and state-level agencies that are expected to increasingly outsource these services..."*).

[19] For example, in the case of a PBL for the energy sector (PE-L1061) the LD proposes a new energy matrix but does not identify the characteristics that this matrix will have. Rather, the PBL identifies a list of activities among a seemingly unconstrained set, without identifying why these activities are the most important activities and what specific bottlenecks they would be addressing. The order in which activities are to be undertaken are likewise not explained. Given the inherently complex nature of such a reform, it is fundamental that the LD contain sufficiently detailed and organized information in order to understand the sector's main problems, the reasons for these problems, or the correctness of proposed solutions.

[20] In the case of the program supporting the mortgage market in Mexico (ME-L1065), LD states that: "The *illiquidity and continued uncertainty in international and local financial markets has translated into greater complications for institutions in the private sector in terms of obtaining funding for mortgage activities, given that direct funding and funding from the secondary market has tapered and become more expensive*." This suggests a counter-cyclical role as the project objective. However, the general objective stated in the LD is not to mitigate the crisis, but rather "to promote the development of efficient and inclusive mortgage markets in Mexico". This inconsistency between a financial emergency intent and development intent is seen throughout the LD. In the case of a CCT (Oportunidades), the LD also incorporates a countercyclical mitigation as an apparent objective of the loan, but does not provide a description of how this mechanism would operate or evidence of its effectiveness as such entails a logical problem. In particular, the LD does not discuss the program's ability to incorporate beneficiaries during a downturn.

[21] For example, the Dominican CCT program (DR-L1039) identifies budgetary restrictions in the supply of services as a central constraint: *"Las restricciones fiscales impiden que en el corto plazo se incremente significativamente el presupuesto dedicado a salud y educación para asegurar que los servicios brindados a los beneficiarios de Solidaridad cuenten con insumos que cumplan con estándares mínimos"*. However, the intervention's payouts tackle the demand side, not the supply side. In the case of Peru's Agricultural Competiveness Program (PE-L1066), there is also a lack of correspondence between the problems identified in the diagnosis and the actions that were formulated. For component 4, the LD identifies as problems the low expenditure in the sector, pervasive weakness in public and private capacities, and the fact that innovation and research activities are neither strategically targeted nor prioritized. However, the proposed solutions do not address these problems adequately, and in particular do not address the issue of adequate funding for innovation.

[22] This is the case of a governance programmatic PBL in Paraguay (PR-L1043), in which the LD mentions that it is necessary to combine short- and long-term policy changes: "*combina[r] acciones de política de corto plazo que serán reforzadas por otras de mediano y largo plazo"*. However, the LD does not detail what these medium-term activities are, how they will work to address issues of efficiency and quality of public spending, and on what timeline they should be implemented.

[23] In the case of Ecuador's costal artesian fishing project (EC-L1059), the LD identifies one component which would promote the competitiveness of the industry, but in doing so may finance activities that could undermine the other component, which would promote the sustainability of the industry. The LD does not provide a discussion of measures which would prevent this from happening. A potential contradiction is also seen in the loan to support service providers for the Mexican petro industry (ME-L1051), whose objective is both to increase competitiveness and to reduce green house gasses (GHG). However, the LD

does not identify any safeguards in the components aimed at competitiveness that would impede this activity from actually increasing GHG.

[24] For instance, in Belize's Agricultural Service program (BL-L1009) the model of intervention is internally inconsistent. Its main rationale is that better final outcomes (increased yield) will follow from the provision of public goods in a new way (competitive grants), supported by further improvements in the phyto-sanitary inspection and regulation system. However, the diagnostic data provided is inconsistent that there is a problem with low yields (indeed, the objective or reducing the yield gap does not make sense since yields are higher than the region). Second, the LD does not provide economic or logical arguments to substantiate the hypothesis that the move from direct funding of research to research supported by competitive grants would benefit producers—beyond the obvious benefit of eliminating taxes on them.

[25] The PDL in the electricity sector of Minas Gerais (BR-L1028) is a good example where the operation is temporally inconsistent. Although the LD wording establishes targets to be achieved in the future, the operation had in fact already achieved the targets, and would only be providing recognition of past expenditures. Given that the logic of a PDL is to disburse based on future results, the retroactive disbursement based on past results is logically inconsistent.

[26] This is the case of the EURUS wind project (LE-1068). The LD mentions that: "*As one of the first private wind power projects in the country, the [Eurus Wind] Project will provide an important demonstration effect*" (Executive Summary). However, the LD does not indicate what is being done so that an isolated operation, which will only produce 0.25 GW out of the 59.5 GW of Mexico's total installed capacity, will create a demonstration effect. Given that the operation is in fact a refinancing of an existing investment, it is even less clear how this financial transaction will produce a demonstration effect.

[27] For instance, in the case of PROCAMPO (ME-L1041), the Bank is financing what amounts to an unconditional transfer of monies to producers, yet the LD does not analyze the mechanisms by which this transfer can be expected to address market failures in agriculture, or how it could lead to improvements in productivity and competitiveness. Evidence of significant liquidity constraints, which would be one mechanism by which an unconditional cash transfer could impact productivity, is not provided, nor is there an explicit attempt to test this hypothesis. Since little is known about the liquidity constraint problem and its impact in productivity, it is unclear whether it can be solved through PROCAMPO in the long run.

[28] This is the case, for example, of the housing program in Surinam (SU-L1015) in which the Bank is financing a second program in the sector. However, the LD does not discuss the lessons learned from the first phase, nor do these form a basis for the second loan. In particular, the first phase documentation emphasized the need for specific reforms in order for the program to be successful. The second loan does not discuss the policy environment, except for brief references to coordination issues between entities.

[29] A good example of this are the "Green Bank" loans, both of which finance parts of the cost associated with Banks creating a "green" headquarters (Mexico and Panama). The LD, however, do not discuss the Bank's financial additionality. Given that these are among the largest financial institutions in their respective countries, it is doubtful that lack of financing was a restriction for the "Green" Headquarters to be built. The LD also does not discuss how the Bank will promote demonstration effects, which would be an additional avenue by which additionality could be furthered. In particular, the LDs do not detail the specific activities that would be financed to disseminate these best practices among other institutions, including institutions which have more limited capacity.

[30] For example, in the supplemental financing projects the LD do not adequately identify the underlying causes of cost overruns, and therefore do not provide evidence that these same factors which produced overruns in the first case would not constitute risks in the future.

[31] For example, in PN-L1056 the LD explicitly recognizes that the B lender will only be an option in the medium-term "*an envisioned B loan (…) targeted at an estimate amount of approximately US$20 million or such other amount to be determined based on market conditions. Given current market conditions, the B Loan would be syndicated to investors in the medium term as market conditions become more favorable. As a consequence, IDB would initially disburse the A Loan, while continuing its efforts to place the B Loan in the market*". However, the LD does not identify the consequences for the project's results, should this B lender not materialize.

[32] In the case of support to Banco Agricola in El Salvador (ES-L1033), the LD does not consider regulatory or normative constraints which could limit the effectiveness of SME credit. In the DR PBL (DR-L1043)

the project does not properly address the political economy of power sector subsidies—both from the point of view of users (who pay very little) and of the providers (who receive large government payouts). The risk that plans to adjust rates and subsidies are not undertaken is identified but the underlying causes of this are not addressed.

[33] In the case of an integrated urban development program (BR-L1171), the LD identifies as risks the possibility that components may not be complementing the Federal Government's Growth Investment Program (PAC); but this risk is endogenous, as the selection of complementary components is within the purview of the project. In other cases LDs identify institutional, fiduciary, or other similar issues as risks, even as these are components of the project (e.g. NI-L1045, GU-L1039, CO-L1079, CO-L1063, BR-L1053, UR-L1058, VE-L1021). In the case of JA-L1027 roads maintenance project the LD identifies sustainability as a risk: "*Once a road asset is rehabilitated there is a risk that it would revert to a state of disrepair before the full useful life of the asset is realized due to a lack of adequate funding of periodic and routine maintenance activities*", but this is precisely what the project aims to correct, so the risk is not adequately defined.

[34] In the case of Fortaleza social infrastructure project (BR-L1122), the project indicates that funding risks may limit the ability of the program to finance all six cultural centers (CUCAs), but the LD then mentions that this has already materialized due to exchange rate appreciation, "*The devaluation of the dollar and the costs of two of the sample's projects(including final engineering designs and the centers' operating and maintenance costs)10 indicate that program resources will be insufficient to achieve these targets*"

[35] In the case of customs reform in Paraguay, the document identifies as a problem the fact that customs does not have funding independence, but then does not treat this risk with mitigation measures. Another instance where risks are inconsistent is in the case of the Jamaica emergency loan. Here the document mentions that "*The Financial risks for the Bank as well as the risks for the DBJ, are low, due to the nature of a second-tier lending operation. The first-tier AFIs will absorb the credit risk from the private enterprises. The AFIs eligible for the lines of credit will provide specific guarantees to the DBJ when the requested amounts are above their allocated credit line. In addition, the operation is backed by a sovereign guarantee.*" But this dismissal of the risks involved is inconsistent with the macro assessment done as part of the loan, which indicates that if multiple shocks occur (as they appear to have occurred due to the financial crisis) the debt situation may deteriorate, "*there are risks to the debt outlook and the sustainability of the debt position will continue to be dependent on the external shocks the country may suffer in the coming years.*" In other cases risks are associated with the inability to properly evaluate projects, but the mitigation is inadequate. The Jamaica citizen security project is a good example, in which the document mentions the lack of a "culture" of evaluation as one reason why there are few results documented, and mitigates the risk by investing in staffing (hiring a statistician). In this case the LD discussion suggests that staffing problems are a consequence of the lack of evaluation "culture" and not the cause.

[36] Lack of indicators capable of measuring aspects of quality (rather than just counts) are seen, for example, in AR-L1092, BA-L1015, BL-L1009, BO-L1047, BO-L1053, BR-L1020, BR-L1021, BR-L1051, BR-L1053, BR-L1078, BR-L1195, CO-L1079, DR-L1043, EC-L1065, GU-L1039, HA-L1028, HA-L1042, HA-L1015, HO-L1033, JA-L1009, NI-L1010, PE-L1057, PR-L1022, PR-L1026, SU-L1015, TT-L1005 and UR-L1058.

[37] This is the case, for instance, of BL-L1009, which include the indicators Capacity Building and Training with target defined as "Cumulative of 60 prospective proposal teams trained 10 research teams". In any part of the document there is a clear explanation of the training to be offered and how to measure its qualities.

[38] Examples of this problem can be found in all non-sovereign guarantee projects , as well as in many sovereign guarantee: BA-L1015, Bo-L1032, BO-L1034, BO-L1043, BR-L1078, BR-L1084, BR-L1171, BR-L1177, BR-L1180, BR-L1181, BR-L1228, CH-L1056, CO-L1028, CO-L1041, CO-L1076, DR-L1030, DR-L1043, GU-L1039, HA-L1039, JA-L1022, ME-L1067, PE-L1057, UR-L1036, UR-L1054 and VE-L1026.

[39] This was observed in CO-L1076. In this project no output indicator was identified, only a list of vaguely described activities, such as contracts for consulting, pre-investment studies and expenditures for support activities. This is true in a large share of projects. In other cases some indicators are included with metric, but completely disconnected from the objective and the diagnosis of the project, such as GU-L1039 that measure the number of women benefiting from the program, when gender was not mentioned as an issue, or NI-L1010, where several indicators relate to the production of coffee, which was not mention anywhere else in the document of a project that aims at supporting the storm water drainage program in Managua.

[40] That is the case of some indicators in the following projects: AR-L1092, BO-L1043, BO-L1047, BR-L1071, DR-L1043, GU-0163-1734, ME-L1065, ME-L1067and NI-L1010. Some projects even have the milestones and targets but they are confusing. For example, AR-L1098 set targets and milestones for some indicators that are identical to the baseline. BA-L1008 has milestones and targets defined as percentage of non-numerical indicators. Many other projects, such as NI-L1010, define the targets (milestones) as percentage changes, but no baseline is provided, raising questions about how such targets (milestones) were computed. GU-L1014 defines as the target of protected areas with cadastral registry established "up to 92", which in principle could be achieved with any number smaller than 92.

[41] For instance CH-L1054, CO-L1041 and CO-L1079 do not include in the Results Matrix some of the indicators described in the text, and therefore it is unclear whether data regarding such indicators will be collected. In the case of BR-L1171, some inconsistencies were found regarding labels and potential indicators listed in the loan document.

[42] Many roads project LDs state as objectives socio-economic development and or competitiveness, but only identify outcome indicators as travel time and operating costs. Another example of this problem can be found in the PROCIDADES projects. They aim at improving quality of life, but then identify indicators that do not measure quality of life, such as housing value indicators, for instance. Regarding a lack of a full set of indicators, there are many examples. For example, the LD for the program to support the electricity sector in Venezuela (CORPOELEC) identifies a set of outcome indicators related to the performance of CORPOELEC, but lacks outcome indicators for the performance of the sector, which is the program's stated objective. Other examples of this are NI-L1010, HO-L1033, HA-L1039.

[43] This is the case of some or all indicators of AR-L1045, AR-L1095, BO-L1047, BR-L1092, BR-L1093, CH-L1033, CO-L1076, DR-L1014, HA-L1040, JA-L1022, ME-L1039, ME-L1067, PN-L1033, PN-L1048, PN-L1049, RG-L1027, TT-L1005, UR-L1036, UR-L1054 and VE-L1026.

[44] For example, BR-L1020 and BR-L1021 set baseline zero when some APLs have already been operating. The same is observed in EC-L1074, which supports the National Program for Social Housing Infrastructure, and the zero baseline is inconsistent with the idea that it is an ongoing program.

[45] Examples of total or partial lack of baselines can be found, for example, in AR-L1095, AR-L1098, BO-L1034, BO-L1047, BR-L1177, BR-L1192, BR-L1193, CH-L1033, CH-L1054, CO-L1041, CO-L1076, CO-L1080, DR-X1003, EC-L1059, EC-L1065, GY-L1021, GY-L1027, HA-L1039, HA-L1042, ME-L1058, ME-L1067, ME-L1084, NI-L1016, PN-L1033, PN-L1048, PR-L1019, PR-L1032, PR-L1043, SU-L1015, TT-L1005, UR-L1016, VE-L1021 and VE-L1026.

[46] This is particularly evident in the PROFISCO projects, for which the baselines were in many cases set in 2005/6 or even earlier. Other examples are most of the Procidades, for which there is no time reference for the baseline, but the loan documents suggest that the data used come from the respective municipality's Master Plan, mostly prepared in 2006 or earlier. Other examples of baselines that are outdated or lack time reference can be found in BR-L1051, CO-L1082, PE-L1057.

[47] For instance, BR-L1177 aims at reducing maternal mortality, which is one of the main indicators identified in the project. However, the baseline in the loan document (paragraph 1.5) is almost twice as big as the one in the results matrix. Similar example can be found in NI-L1010, where the baseline of the indicator of number of inhabitants at risk is 676 in the annex and 600 in the loan document. No explanation for such inconsistencies was found in the projects.

[48] Lack of budget for monitoring and evaluation, or at least not being clearly specified in the project was a common fact in 2009. For examples see: BL-L1009, BO-L1031, BR-L1051, CO-L1028, CO-L1063, CO-L1080, DR-L1043, GY-L1027, ME-L1065, NI-L1023 and PE-L1040.

[49] Many examples can be mentioned in this case. For instance, BR-L1020 and BR-L1021 do not design any impact evaluation under the argument that the evaluation of the project should be done by OVE. NI-L1016 states that impact evaluation will be done only in the case resources are available at the end of the project, which suggest lack of planning the Monitoring and Evaluation System.

[50] GU-L1039 claims that "*As part of the evaluation strategy, an impact evaluation based on an experimental design will be performed*", but neither the loan document nor the monitoring and evaluation annex mention a control group, for example. The same is true for HA-L1042. NI-0155 mentions the use of experimental design, but for only two indicators the use of control groups is mentioned, although the nature of the control groups is not clarified in any part of the project. That is similar to JA-L1009, which also reveals an intention of doing "*rigorous impact evaluation*", but the selection criteria for the control communities is not clear in the document.

[51] Also, although a smaller share of SG operations scored adequate in diagnostic, objectives, and risks, among those that scored inadequate, scores were concentrated in the 2 category, whereas for NSG operations they were concentrated in the 1 category.

[52] In the case of supplementary financing operations, the low scores were due to (i) a partial identification of causes of cost overruns, (ii) an inadequate calculation of rates of return with and without the supplementary financing, (iii) no updating of results frameworks to reflect results achieved thus far and to reflect additional benefits identified, (iv) a risk analysis that does not treat other possible causes of cost over-runs. As can be seen from the table, this is reflected in low scores throughout, but particularly in the identification of results, baselines, and monitoring framework. In the case of emergency lending, the main issues were (i) improper diagnostic of the effects of the global downturn on both the supply and the demand for credit by both financial institutions and by firms, (ii) an inconsistency between the stated intent of operations—which was to restore credit to firms—and components, which were aimed at addressing financial commitments of Banks and borrowers, and (iii) the absence of a results framework.

[53] The low scores of the programmatic PBLs were associated with (i) a poorly identified problem situation, particularly given that programmatic PBLs usually deal with structural changes in the sector; (ii) results framework that focused on specific activities to be undertaken and not on the expected results of these activities; (iii) a poor explanation of the role of the different tranches of the programmatic in achieving reform objectives. Thus, only in one case (Peru's water sector PBL—PE-L1040) did a PBL obtain an adequate rating in a substantive dimension. In the remaining 15 cases the ratings were of a `1' or `2' for all operations.

[54] Instruments that are rarely used were not included in the table. There are "ERF","FAB","FAC","IGR", "TCR","PDL", and "INO" modalities.

[55] Minutes of the 20 February 2003 meeting Policy and Evaluation Committee (PEA/03/3).

[56] RE-333-4 Analysis of project evaluability – Year 2005. Report of the Chairman of the Policy and Evaluation Committee

[57]

Issues included in the Norm (PR-1002)	
Links to country Strategy	38.10%
Technical Alternatives	21.43%
Environmental Strategy	69.05%
Risk Analysis and Mitigation Measures	64.29%

[58] At the discretion of the VPS Division chief a QRR meeting may be convened to discuss either the document as a whole or specific policy and institutional issues.

[59] The incidence with which evaluability considerations are brought up is, on the whole, similar to the findings reported by this office for CRGs in 2005.

[60] The only exception being VPS/ESG that with some discipline expressed the unit position vis-a-vis the projects, applied the standards and protocols as established in the social and environmental safeguard policy.

[61] These findings for projects reviewed and approved in 2009 are consistent with the interim assessment of the NPC carried out in September 2008 by the Bank (OP-140). This assessment pointed out that project teams are now mostly flying solo or unsupervised. Despite minor modifications of processes arising from this assessment, the QRR was essentially ill-prepared in 2009 to perform as a quality review mechanism.

[62] Considering extremely low participation in QRR, interdepartmental interaction occurs and it is recorded only available at the team-level. No medium-level nor high-level responsibilities of check and balances are provided at the QRR level.

[63] SPD provided OVE with files and documentation regarding the DEM for analysis. The files document the different iterations of the DEM matrix, as it was reviewed by the project team and by SPD. All projects, with the exception of HA-L1015, had a DEM matrix consistent with that sent to the Board.

[64] The TRS category also includes time spent on the retrofitting of DEM checklists to 2008 projects, time spent on the production of the Development Effectiveness Overview, and time spent on other activities related to the revisions of the DEM instrument itself.

[65] Comments received by SDP indicate that actual staff time amounted to an average of 11 hours per project if evaluability hours reported under the project codes are included. OVE cannot validate this, given that project TRS codes do not identify a separate sub-code for evaluability.

[66] The meetings were called by different actors: VPS - SCF (BR-L1192-BR-L1193), VPC (EC-L1074, BR-L1180, AR-L1098), EVP (BR-L1260), and LEG (PR-L1043).

[67] It should be noted that evaluability requirements are not considered a bank policy, and therefore problems in evaluability do not constitute a basis, according to the normative, for consideration by OPC.

[68] For an analysis of the checklist approach see Atul Gawande, The Checklist Manifesto, (2009)

[69] In theory, more demanding yes/no questions could be designed, such as asking: "Is all of the empirical data needed to describe the problem accurately present?" Such questions, however, generally demand a qualified and nuanced response, not a simple yes or no.

[70] RE-333-4 Analysis of project evaluability – Year 2005. Report of the Chairman of the Policy and Evaluation Committee