ove Office of Evaluation and Oversight

Project Evaluation

2011 Evaluability
Review of Bank Projects

IDB

Inter-American Development Bank

June 2012

Abstract

During the last decade, the IDB has strongly promoted an agenda to increase its capacity to report on results, with a particular focus on project evaluability. The first pilot exercise to validate the Development Effectiveness Matrix (DEM) of randomly selected projects began in October 2011. The objective was to provide suggestions for improving the DEM as a tool for assessing evaluability.

In validating the DEM for Sovereign-Guaranteed (SG) loans, results indicate significant progress in SG project design, although there is still room for improvement in a number of areas. The Non-Sovereign Guaranteed (NSG) DEM was not validated; instead, the evaluability dimensions of five NSG projects were analyzed, revealing that NSG projects are relatively weak in evaluability.

In light of the findings, the DEM for SG operations should be strengthened; the DEM for NSG operations should be thoroughly revised; there should be one unit in IDB overseeing evaluability for all projects to ensure consistency and comparability of treatment; and post-approval follow-up to monitor and assess results should be ensured.

Original: English

2011 Evaluability Review

of Bank Projects

Office of Evaluation and Oversight, OVE

Inter-American Development Bank

June 2012

Office of Evaluation and Overishgt
1350 New York Avenue, N.W.
Washington, D.C. 20577
www.iadb.org/ove

Content

Acronyms

CBA	Cost Benefit Analysis
CEA	Cost Effectiveness Analysis
DEF	Development Effectiveness Framework
DEM	Development Effectiveness Matrix
DEO	Development Effectiveness Overview
ECG	Evaluation Cooperation Group
ERM	Eligibility and Risk Review Meeting
EVP	Office of the Executive Vice President
GPS	Good Practice Standards
LD	Loan Document
LRR	Loan Results Report
NSG	Non-Sovereign Guarantee
NPC	New Project Cycle
OVE	Office of Evaluation and Oversight
PBL	Policy Based Loans
PBP	Programmatic Policy Based-Loans
POD	Project Operation Document
QRR	Quality and Risk Review
SG	Sovereign Guarantee
SPD	Office of Strategic Planning and Development Effectiveness
VPP	Vice-Presidency for Private Sector

The concept of evaluability is central to the Bank's results agenda, as it is impossible to determine the results achieved by IDB projects in Latin America and the Caribbean unless they are designed up-front to be evaluable downstream.

(C) E. Virgen, 2007

Executive Summary

Assessing evaluability means determining ex-ante whether a project will be able to be evaluated when it is completed. The concept of evaluability is central to the Bank's results agenda, as it is impossible to determine the results achieved by IDB projects unless they are designed up-front to be evaluable downstream.

During the last decade, the Inter-American Development Bank (IDB) has strongly promoted an agenda to increase its capacity to report on results, with a particular focus on project evaluability. OVE has periodically reported to the Board of Executive Directors on project evaluability since 2001. In 2008 Management designed the Development Effectiveness Matrix (DEM) to be included in all IDB projects submitted to the Board for approval. Two different types of DEMs were developed: one for Non-Sovereign Guaranteed (NSG) loans and one for Sovereign-Guaranteed (SG) loans. The SG DEM included an evaluability self-assessment tool, while the NSG DEM was designed to be a development results assessment tool. In early 2011 Management further changed the format and revised the process for the DEM.

This report has two objectives: (1) to report on the findings of OVE's first (pilot) validation round of project evaluability and (2) to provide suggestions for strengthening the DEM as a tool for assessing evaluability. In 2011 OVE adopted a new process of reporting on project evaluability by directly validating a random sample of Management's DEM scores. A sample composed of 16 SG projects and 5 NSG projects – approximately one-third of the projects approved under the new DEM format and procedure -- was analyzed during this exercise. When approved by the Board, Management's reported evaluability scores for the SG projects in the sample ranged from 6.3 to 10 (out of a maximum of 10), while the scores for the NSG projects ranged from 8.3 to 10.

In validating the DEMs for SG loans, OVE's final scores were similar to those of Management, though somewhat lower on criteria related to problem diagnosis and project logic. This is a positive result that indicates significant progress in SG project design. As a note of caution, the report presents some evidence that even SG projects with high DEM scores may not ultimately be evaluable if project teams do not have adequate incentive to follow up on monitoring and evaluation needs post-approval. Specifically, OVE's review of post-approval Loan Contracts and Loan Results Reports for the projects reviewed do not indicate that evaluability aspects missing at approval were later addressed as intended.

OVE's analysis also suggests that the design of the SG DEM can be further improved in a number of areas. These are outlined in OVE's recommendations below.

OVE was unable to validate the NSG DEM and instead opted to independently analyze the evaluability dimensions of the five NSG projects. As currently designed the NSG DEM tries to fulfill multiple roles, as it seeks to predict the likely achievements of the project, identify monitoring indicators, and assess evaluability. As a result, the evaluability analysis is constrained and does not properly measure evaluability according to the standards agreed to in Management's 2011 DEM review (GN-2489-4). The five NSG projects reviewed by OVE were considered to be relatively weak in evaluability, though the NSG DEM tool indicated very high evaluability scores.

RECOMMENDATIONS

OVE makes the following recommendations for Management to further strengthen the DEM tools and their application:

- **Strengthen the DEM for SG operations by:**
  - **Providing more detailed guidance for questions with greater subjective content.** This was particularly true for the first six questions of the DEM, whose tutorial need to be clarified and questions need to be refined.
  - **"Nesting" the questions on indicators rather than answering each question independently.** The DEM asks whether at least one indicator for each impact/outcome/output has been identified, whether indicators are SMART (specific, measurable, attributable/achievable, relevant and time-bound), and whether they have baselines, targets and means of verification. All these questions should be considered in a group.
  - **Reviewing the criteria on external and internal validity.** They need to be considered together, as external validity is related to the extent to which internally valid results will be held to be true for other interventions in

similar cases. Also, the tutorial of these criteria needs to consider how to handle innovative projects. A pioneering project, which defines well the model to be tested and how to do so, may not be less evaluable than a project which is replicating what other interventions have done in the past.

- **Eliminating the requirement to extend the results framework and indicators to ultimate impacts.** Although, form the point of view of vertical logic it is important to identify impacts and relate them to the outcomes, projects should not be required to identify indicators to measure impacts that go beyond the intervention.

- **Taking the Risk Management Section score out of the overall evaluability assessment and scoring.** Although the risk section is important and should be kept in the DEM for the purpose of assessing project quality, its format is not adequate for measuring evaluability, and the typically high scores tend to bias the overall evaluability scores of projects.

- **Revising the criteria for PBLs and PBPs in three main aspects.**

  - **Rethink the required economic analysis.** Currently, Economic Analysis section is computed as the maximum between the CBA and the CEA. Yet neither a CBA nor a CEA is applicable to PBLs and PBPs. An alternative would be to specify the score of this section as being the maximum score among the CBA, CEA or general economic analysis in the case of PBLs/PBPs. This general analysis would be expected to provide evidence on the likely economic impacts of the policy and institutional reforms supported by the PBL/PBP.

  - **Clarify the standards for output indicators.** In some cases the conditionalities were accepted as output indicators and in others they were not. The standards for accepting or rejecting them as indicators should be made clear.

  - **Clarify the unit of analysis.** In particular, the unit of analysis should be well defined in the evaluability standards to clarify whether the DEM should consider the program as a whole or individual tranche of a PBL separately and independently. The same is true for projects that are expected to be complemented by an investment loan.

- **Thoroughly revise the DEM for NSG operations and improve their evaluability by:**

  - **Disconnecting the assessment of evaluability from the assessment of likely development impact.** The dual purposes of the current DEM constrain the tool's format and limit the assessment of project evaluability. OVE recommends that IDB develop a more comprehensive evaluability assessment tool, free from the constraints imposed by the current DEM structure.

  - **Ensuring that the revised NSG evaluability tool covers the essential aspects of evaluability.** Evaluability is a straightforward concept that should be applied equally to any kind of intervention. Its assessment has to be grounded in at least three dimensions: project logic, including the diagnosis of the market failure the project is trying to address and clear definition of the objectives the Bank aim to achieve with such intervention; economic analysis, which should attempt to measure all the impacts the intervention will have on the targeted market through a proper analysis of supply and demand; monitoring and evaluation, precisely setting the plan for who is going to be responsible for collecting information and how the information will be used to measure the impact of the intervention.

  - **Including a Results Matrix to clearly present all output and outcome indicators and the links between them.** A Results Matrix would help organize relevant information and show project logic and output and expected outcomes in one place (rather than separately as forced by the current DEM). The NSG DEM Guidelines mandate that projects include an Output Table, which could fulfill this role. However, in practice such table does not appear to be in use.

  - **Clearly identify funding for project evaluation prior to approval.** Funding may come from NSG loan proceeds (financed by borrower) or, if this is not feasible, from alternative financing sources, but it needs to be identified up-front.

- **Consider having one unit in IDB oversee evaluability for all projects -whether SG or NSG- to ensure consistency and comparability of treatment.** One option is that suggested by Management in GN-2489-4: "In 2012 a review of the effectiveness of this process will be completed and if appropriate SPD will take this responsibility and will also begin to validate the ratings for the NSG DEMs." The first recommendation above on DEM content suggests that the SG and NSG DEMs need not be as different as they currently are. Bringing responsibility for both under one roof might help eliminate unnecessary inconsistency and promote learning and innovation that serves both SG and NSG clients.

- **Ensure post-approval follow-up to monitor and assess results downstream.** This report raised an empirical question regarding the incentive structure, which seems to work against a culture of reporting on results. OVE will further address this issue in subsequent evaluations.

1

During the last decade, the IDB has strongly promoted an agenda to increase its capacity to report on results, with a particular focus on project evaluability.
(C) Guerrero, 2012

1 Introduction

Evaluation is an important tool for any organization to promote learning and improve its effectiveness. Assessing evaluability is a way to determine ex-ante whether a project will be able to be evaluated when it is completed. Very often the capacity of the evaluator to do a proper assessment of a project's progress and downstream results is hindered by its design.

For instance, lack of baselines or adequate indicators and mechanisms for data collection can undermine the ability to learn from ongoing projects. The literature has shown that "*Evaluability assessment is […] cost-effective because it can prevent costly evaluation of programs and practices when the logic models are not plausible or when the programs still need to develop better and more relevant activities.*"[1]

During the last decade, the Inter-American Development Bank (IDB) has strongly promoted an agenda to increase its capacity to report on results, with a particular focus on project evaluability. In 2001, the Office of Evaluation and Oversight (OVE) developed tools to assess project evaluability[2] and applied them to all projects approved throughout the year. This exercise was repeated in 2005 and 2009.[3] In 2008 the Board of Executive Directors (the Board) approved a new Development Effectiveness Framework (DEF), and Management designed the Development Effectiveness Matrix (DEM) to be incorporated into all IDB projects before being submitted to the Board for approval. Two different types of DEMs were developed: the Non Sovereign Guaranteed (NSG) and the Sovereign Guaranteed (SG).[4] The SG DEM included an evaluability self-assessment tool, while the NSG DEM was designed to be a development results assessment tool. OVE's 2009 evaluability exercise was performed parallel to the implementation of the DEM. In the beginning of 2011, OVE and Management worked together on the harmonization of the criteria to assess evaluability, and Management proposed a new DEM and a revised process.[5]

Management's latest review of the DEM (GN-2489-4) sets the evaluability standards to be followed by SG and NSG operations in the IDB[6] (Box 1). Both substantive and procedural aspects are addressed. Substantively, the first set should include an analysis of the problem to be addressed and its context, the definition of the objectives to be achieved, the logic behind the intervention chosen and the consideration of risks the operation might face and their respective mitigation measures. Procedural standards are meant to guarantee measurability of the results and progress of the intervention, including the definition of output and outcome indicators, their respective baselines and targets, and a monitoring and evaluation system. In the case of NSG operations, projects have to have a clear discussion regarding the market failure being addressed and the additionality of the Bank.

In response to the development of the DEM, OVE has recently adopted a new approach to overseeing project evaluability. The 2010 agreement for the Ninth General Increase of the Resources of the Bank (IDB-9) requires that OVE report annually to the Board on project evaluability and establishes the DEM as "*a common instrument for rating evaluability of projects at entry." Rather than continuing to use a wholly separate system to review project evaluability, as was done in 2001, 2005, and 2009, OVE has adopted a system of continuous validation, on a sample basis, of DEM scores produced by the self-evaluation system.*[7]

Box 1. Evaluability Standards

In 2011 Management conceptualized the standards through which evaluability should be assessed, and they are expected to "be met by Sovereign Guaranteed operations (SGO) and Non-Sovereign Guaranteed (NSG) operations at the IDB." (GN-2489-4). According to such standards six dimensions should be considered: Diagnosis, Definition of Objectives, Logic, Assumptions and Risks, Indicators, and Monitoring and Evaluation.

Diagnosis: The project document should include an analysis of the problem the project aims at solving, or in the case of the private sector loans, the market failure or need of the client to be addressed. It should be data-driven and evidence-based to provide enough information on the nature and scope of the problem to be addressed and the role of the proposed intervention in addressing such problems.

Definition of Objectives: "Project objectives define the project's reason for being" and should be a specific statement of intent, identifying the development results expected from the intervention.

Logic of the Project: "The logic of the project must elucidate how its objectives are consistent with the proposed solution to the problematic presented in the diagnosis, and how its components are logically related and necessary to the achievement of the purpose."

Assumptions and Risks: The project should identify risks that could directly affect the execution of the project and the attainment of its goals, and any mitigation measure to be taken if the identified risks materialize.

Indicators: The project should identify output and outcome indicators which are specific, measurable, achievable/ at-tributable, relevant and time-bound. Such indicators should have clearly defined baselines, targets and means of verification.

Monitoring and Evaluation: The project should define the means through which the project will be monitored and eval-uated, including the budget, means for data collection, responsibilities and method to be used to track and estimate project results.

This report has two objectives: (1) to report on the findings of the first (pilot) validation round of project evaluability, and (2) to provide suggestions for strengthening the DEM as a tool for assessing evaluability. The first goal responds to the IDB-9 mandate to report annually on evaluability, while the second responds to the Board's statement in early 2011 that "*Directors agreed that the new DEM was a "work in progress" which would require adjustments as it is being implemented. (…). It was also agreed that OVE should evaluate its application at the end of the first year.*"[8]

2

Project teams prepare the DEM, and SPD validates the final scores. In October 2011, OVE began its first pilot exercise of validating the DEMs of randomly selected projects.

2 The DEM and OVE's Validation System

A. DEM Criteria and Process

The criteria and format of the SG DEM have been reviewed and improved since 2009, increasing its capacity to assess evaluability.[9] The SG DEM was developed by SPD as a "review checklist" divided into three sections: (i) Strategic Alignment, (ii) Development Outcome – Evaluability, and (iii) IDB's Role – Additionality.

The evaluability section of the DEM is detailed in Box 2 and Annex A and is the only part of the tool considered in this report. It has binary criteria to which a YES (1) or NO (0) should be assigned. DEM scores are a weighted sum of the scores on each criterion and range from zero to ten. Project teams prepare the DEM, and SPD validates the final scores.

Box 2. Evaluability in the SG DEM

The second section of the DEM, which assesses evaluability, has four sections: (1) Project Logic; (2) Economic Analysis; (3) Monitoring and Evaluation and (4) Risk Management. Each section sums to ten and has equal weight in the final evaluability score.

Project Logic Section: This section has 33 checklist questions in three subsections: Program Diagnosis, Proposed Interventions or Solutions, and Results Matrix Quality. The purpose of this section is to analyze whether the Loan Document (LD) has a proper diagnosis, including identification of the development problems to be addressed, their root-causes, and how the intervention will address them. It also assesses whether the LD provides enough evidence of the effectiveness of the proposed intervention and a clear vertical logic (i.e. how inputs and activities will be translated into outputs, which will lead to outcomes, which will contribute to final impact). The Results Matrix sub-section asks about identification of SMART (Specific, Measurable, Attributable/Achievable, Relevant and Time-Bound) indicators, including baselines, targets and means of verification.

Economic Analysis: According to the criteria in this section, projects should present either a cost-benefit analysis (CBA) or a cost-effectiveness analysis (CEA). These should assess the economic benefits and costs of the intervention and potential alternatives. All assumptions should be explicitly mentioned and supported by evidence, and a sensitivity analysis should be included. If neither a CBA nor a CBE is included in the project, a proper justification should be presented.

Monitoring and Evaluation: All projects should comply with a minimum standard for monitoring and evaluation. The DEM includes 16 questions in this section regarding the necessary *"elements for projects to include systems to track implementation and measure results"* (GN-2489-4), such as definition of monitoring and evaluation plans, budget for the activities, and methods.

Risk Management: This section has four compliance questions related to the Risk Matrix. The first two questions ask whether the likelihood and magnitude of identified risks are measured. The last two questions relate to proposed mitigation steps and indicators for monitoring their implementation. All questions have the same weight.

Although all questions are binary, some require more judgment by the evaluator than others, such as: *"The main problem is clearly identified,"* or *"the main factors (or causes) contributing to the problem are clearly identified,"* or, following the tutorial for specific country characteristics required in question 6, *"(…) the POD includes, in a precise manner, the historical, political, social, and economic context in which the problem arose, as well as those in which its solution will be attempted."*

The NSG DEM is very different from the SG DEM. It has also gone through changes since its implementation, especially to include evaluability, which was not part of its initial design. The NSG DEM serves two purposes. First, it justifies an IDB intervention by rating the project's potential development outcomes, its potential contribution to IDB/Country's strategic priorities, and IDB's additionality. Second, it attempts to assess project evaluability based on a review of its Development Outcome Section.[10] An evaluability score is given to each indicator, ranging from zero to ten, derived from an assessment of (i) the quality of indicators/results, (ii) the rationale of the project score, (iii) the analysis/evidence included, and (iv) the monitoring and evaluation arrangement. Similar to the process for SG projects, teams are initially responsible for filling in the DEM, and the ratings are later validated by the VPP Development Effectiveness Officer. Box 3 explains the details of how evaluability is considered in the NSG DEM.

Box 3. Evaluability in the NSG DEM

According to the NSG DEM Guidelines, the NSG evaluability instrument was designed to fit the nature of private sector operations, which have a different logic from that for the SG projects -- namely to address a specific market failure or a business need of the client (rather than a country's development problem), with positive development outcomes as an externality of this intervention. It is built on the Evaluation Cooperation Group Good Practice Standards (ECG-GPS) for private sector evaluation, with the assumption that the evaluability of an NSG operation should be assessed in terms of whether and how the Loan Document and the DEM worksheet facilitate its eventual evaluation. Evaluability is considered in three Sections of the DEM--Business Performance, Development Outcome, and Private Sector Development, and the indicators for each are to be judged according to four dimensions:

Quality of the indicator/ result: "*Either quantitative indicators or qualitative benchmarks should be included. If quantitative indicators are included, these indicators should meet the SMART (specific, measurable, achievable, relevant and time-bound) indicator criteria. If qualitative benchmarks are included, they should also be SMART indicators. Particularly, definition, target, baseline, and timeline need to be specified."* (NSG DEM Guidelines)

Rationale of the scoring: "*an effective rationale of the scoring [following the ECG-GPS] should be included in the DEM Worksheet."* (NSG DEM Guidelines)

Diagnostics and analysis: "*diagnostics, analysis and/or relevant information should be presented in DEM Worksheet to support the reasonableness of the expected results and to justify the DEM scores."* (NSG DEM Guidelines)

Monitoring and Evaluation: "*Monitoring will be carried out for the area indicators and benchmarks, when a time frame is specified. Otherwise, the results will be verified at the time of self-evaluation. In the DEM Worksheet, the timing of tracking information and the party responsible for collecting the information should be specified."* (NSG DEM Guidelines)

A comprehensive Evaluability score is calculated based on weights assigned to each indicator, and a DEM Worksheet annexed in all projects contains detailed information supporting the ratings.

As noted earlier, in 2011 the process of preparing the DEM was also revised to give teams more guidance. According to the Revision of the DEM, "*Management will implement the changes presented in this document to all projects that will hold their QRR meetings on or after April 1, 2011."* Project teams are now expected to receive comments on evaluability, either from SPD or from the VPP Development Effectiveness Officers, at the Eligibility and Risk Review Meeting (ERM) and the Quality and Risk Review (QRR). LDs should also have an evaluability note along with the validated DEM scores. In practice, all projects that went to the Board after April, 2011 had the new DEM template, but not all of them had the new process of preparing the DEM fully implemented. Most projects with the new process started to be sent for approval only after September, 2011. OVE restricted its validation to those projects.

B. OVE's Validation System

In October 2011, OVE began its first pilot exercise of validating the DEMs of randomly selected projects. Projects were selected as described below, and the LD and relevant annexes were distributed to evaluability teams, who were responsible for reviewing Management's self-assessments and confirming or revising the scores following the same criteria used by project teams.[11] OVE's entire staff and consultants was involved in this exercise, which was overseen by an evaluability panel composed of OVE's Director, three OVE Staff members, and a Research Assistant. OVE evaluability teams were asked to both assess evaluability and indicate issues with the tools being used.[12] After all SG projects in the sample were reviewed, the validation forms prepared by the evaluability teams were sent to SPD for review and comment. (A different process was required in the case of NSG projects, as explained in Chapter 4.)

The sample for the pilot was composed primarily of projects approved after September, 2011, when the implementation of the new DEM process had been completed. Projects were randomly selected with a probability of one-third. A total of 21 projects were reviewed – 16 SG projects and 5 NSG projects. Table 1 in the Annex B lists the 21 projects and their evaluability scores at approval. Table 1 shows the proportion of projects by type of instrument (for SG) or unit (for NSG). The relatively large proportion of PBPs in the sample is due to the higher incidence of this type of project later in the year[13].

Table 1. Percentage of Projects by Instrument

Instrument	OVE's Sample	All Projects approved in 2011
Sovereign Guarantee	76.2%	79.2%
Investment	52.4%	63.0%
PBL, PBLs and EME	23.8%	16.2%
Non-Sovereign Guarantee	23.8%	20.8%
of which OMJ	9.5%	7.1%
of which SCF	14.3%	13.6%

Note: OVE's sample only included PBPs and one PBG.

The distribution of Management's self-reported evaluability scores in the validation sample is similar to the data presented in the 2011 DEO for all 2011 approvals (Figures 1 and 2). A higher variability in scores among SG projects (6.3-10) is observed vis-à-vis the NSG DEM scores (8.3-10). The fact that the validation sample has somewhat higher scores than the overall average for 2011 among the SG projects may reflect the fact that OVE chose only projects subject to the new DEM rules, for which SPD support was stronger. As argued in the 2011 DEO, this close help from SPD should lead to more evaluable projects. Under this assumption, the distribution of scores for the sample and the total projects approved under the new DEM process should be more aligned.

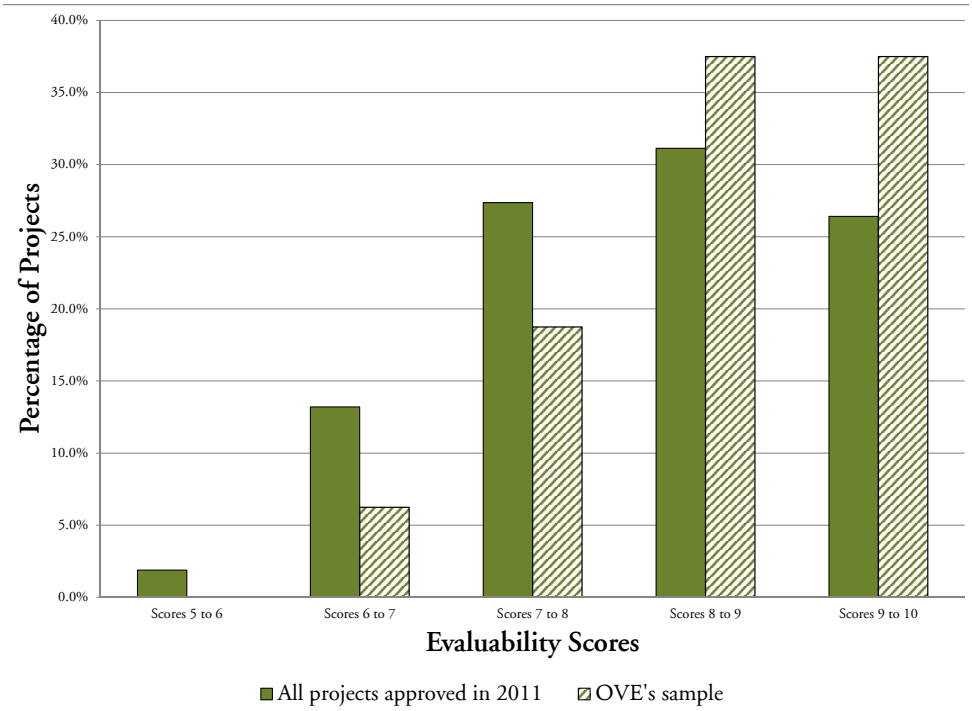


Figure 1
Distribution of Management's Reported SG DEM Scores

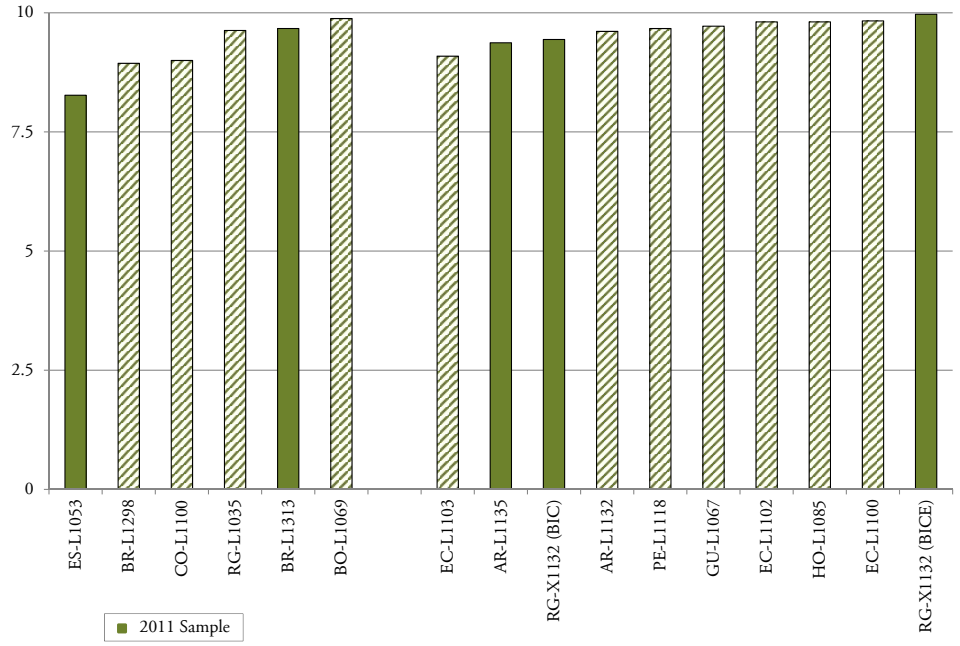


Figure 2
Management Reported NSG DEM Scores
Note: All projects with new template and process were included.

3

3 Sovereign-Guaranteed (SG) Projects

A. Validation Findings: Evaluability at Project Approval

The results of OVE's validation exercise by project are summarized in Table 2. OVE's ratings are quite similar to Management's ratings, with slightly lower OVE scores in most but not all projects. Most differences are concentrated in the first section, Program Logic, where more subjective criteria are concentrated (see Box 1). OVE's average Program Logic score of 7.2 for the 16 projects in the sample compares with Management's average score of 8.2.

Two projects, BH-L1028 to support the New Providence Water Supply and Sanitation Systems upgrade and EC-L1099 for the National Urban Development Program,[14] score under the threshold in the Project Logic Section,[15] according to OVE's validation.

Many of the evaluability problems identified in the SG projects are related to lack of data to support the diagnosis presented in the LD. Many projects identify the problem and list the potential root-causes but do not include a full discussion, based on evidence, to relate the root-causes to the problems being addressed. OVE identified weaknesses in empirical evidence for the main determinants and magnitude of the deficiencies in 13 out of the 16 projects. For example, in BO-L1063, the program for improving municipal management, the LD identified the problem as low tax revenue and suggested in the LD that it is caused by an outdated cadastre system. However, it is not clear that the reason for low tax revenue is the outdated cadastre. In fact, given the lack of information on the determinants of the problem, one could argue the possibility of inverse causality: the government of Bolivia did not invest in cadastre because it could not tax residents.

The Project Logic score is also affected in many cases by lack of internal validity and vertical logic. Proper evidence of the effectiveness of the proposed interventions

Table 2. Total Evaluability Scores and Scores by Section

PROJECT	Project Logic Section		Economic Analysis Section		Monitoring and Evaluation Section		Risk Management Section		Evaluability	
	DEM	OVE	DEM	OVE	DEM	OVE	DEM	OVE	DEM	OVE
AR-L1124	6.1	6.5	6.4	5.2	6.3	6.3	10.0	10.0	7.2	7.0
BA-L1021	8.3	8.1	10.0	8.5	6.6	6.6	10.0	10.0	8.7	8.3
BH-L1028	7.2	3.4	10.0	10.0	7.1	7.1	10.0	10.0	8.6	7.6
BO-L1063	8.2	6.0	7.0	5.5	6.6	6.6	10.0	10.0	8.0	7.0
BO-L1065	8.4	7.1	10.0	10.0	10.0	10.0	10.0	10.0	9.6	9.3
BR-L1282	9.6	8.7	7.0	7.0	6.4	6.4	10.0	10.0	8.3	8.0
CO-L1103	10.0	6.8	10.0	10.0	6.5	6.5	10.0	5.0	9.1	7.1
DR-L1035	10.0	10.0	10.0	8.5	9.1	9.1	10.0	10.0	9.8	9.4
EC-L1087	9.3	8.0	8.5	10.0	6.7	7.8	7.5	2.5	8.0	7.1
EC-L1098	6.3	8.0	10.0	10.0	8.7	8.7	10.0	10.0	8.8	9.2
EC-L1099	7.1	4.7	6.4	7.6	9.1	8.6	10.0	10.0	8.1	7.7
HA-L1065	6.5	7.2	8.5	7.0	5.1	5.1	5.0	5.0	6.3	6.1
HO-L1079	6.9	6.4	10.0	10.0	7.5	7.8	10.0	10.0	8.6	8.6
PN-L1070	10.0	7.5	10.0	10.0	6.5	6.5	10.0	10.0	9.1	8.5
PR-L1061	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
PR-L1066	7.6	6.4	10.0	10.0	10.0	10.0	10.0	10.0	9.4	9.1
Average	8.2	7.2	9.0	8.7	7.6	7.7	9.5	8.9	8.6	8.1

was missing in 12 out of the 16 projects reviewed. In some of these cases OVE judged the evidence to be unsuitably articulated, as it is the case of PR-L1066, Program to Support Job Placement. The project has an extensive literature review, but some of the academic articles discussed seem to suggest that the proposed intervention is not cost-efficient. The lack of evidence on effectiveness also affects the vertical logic of the projects, as in many cases the necessary evidence to support the link between outputs and outcomes/impacts was not found in the documents.

While all 16 projects complied with the requirement in the new DEM to have an economic analysis, the quality of such analyses varied. Two cases - AR-L1024, which aims at strengthening of the Ministry of Economy and Finance, and BO-L1063, for improving municipal management - had weak analyses, the former incomplete and based on unrealistic and unproven assumptions and the latter lacking proper assessment of economic benefits and outcomes, clearly-spelled out assumptions, and sensitivity analysis. In some projects the CBA was not properly reflected in the project, as the assumptions about the benefits did not match the indicators in the Results Matrix. For example, EC-L1098, Program on Citizen Security, has a very good CBA in the annexes, but the Results Matrix did not include targets for results. Poor analyses were also observed in the case of the PBPs, which were required to present a CBA or a CEA even though such analyses are not applicable. This issue is discussed in the next section.

Table 3. Percentage of Types of Evaluation Proposed

Method proposed to evaluate results	Percentage
Random Assignment	25.0%
Non-Experimental Methods	25.0%
Ex-post Cost-Benefit Analysis	56.3%
Ex-post Cost-Effectiveness Analysis	6.3%
Before-After or With-Without Comparison (no attribution)	37.5%

Note: Projects can have more than one method proposed, in most of the cases to evaluate different components.

Although most projects had some plan for monitoring and evaluation, the maturity of such plans varied widely across projects (Table 3). Almost all projects envisioned at least an ex-post cost-benefit (or cost-effectiveness) analysis. Out of the 16 projects, seven included a plan for an impact evaluation of at least part of the intervention, either through random assignment or quasi-experimental methods. Two planned to simply follow the Results Matrix and present a before and after analysis. Regarding the maturity of the plan, two contrasting cases of impact evaluations were PR-L1061, Program on Cadastre and Property Registry, and EC-L1098, Program on Citizen Security. While the first project contained a well-elaborated evaluation proposed, the latter, which received an equally high score in this DEM section, only noted that different evaluative questions would be investigated through different methodologies, without further detail in the Monitoring and Evaluation Annex. In fact, the Annex suggested that impacts beyond the time-frame of the intervention would be analyzed, raising doubts about whether a serious evaluation was contemplated.

The Risk Management Section of the DEM primarily addresses compliance of the Risk Matrix, and OVE's validation results agreed with Management that almost all projects fulfill all requirements. Thirteen out of the 16 projects scored ten in the Risk Management section, with the ones scoring less than 10 lacking proper mitigation measures and indicators to track them. However, it is important to highlight that "*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" (RE-379). In OVE's view a proper risk analysis should go beyond the assessment of the risk matrix, which is limited to identified risks (see Section C below).*

B. Follow-up on Evaluability Issues Post-Approval

The problems noted above have been identified by OVE in prior reviews, particularly those related to incomplete diagnosis and baseline identification or vague identification of the evaluation method. In response to OVE's 2009 review, project teams argued that the absent information necessary to improve evaluability would be collected or clarified during execution as allowed by the New Project Cycle (NPC). The NPC requires project teams to prepare a Loan Results Report (LRR) 18 months after the first disbursement. This document should provide information "*on the completion of any activities which required further technical work after Board approval" and "on activities that have improved evaluability based on the development effectiveness matrix (DEM) at entry." Therefore, the LRRs could address some of the evaluability issues identified at approval.*

OVE's analysis of LRRs suggests that the instrument has not been used as proposed. In many cases LRRs have not been prepared at all, and those that have been prepared have not provided the information missing at approval. Since the implementation of the NPC, 49 projects should have prepared an LRR as of December 2011. OVE found 46 LRRs produced in this time frame, but only 25 were LRRs for projects required to have one, which means that only approximately 51% of the 49 projects have complied with the requirement.[16] The documents prepared for projects approved in 2009 were reviewed by OVE to see whether evaluability issues raised in the 2009 evaluability assessment were treated in the LRR. Of six projects with LRRs, only one included the baseline that had not been computed at the time of approval (provided in the PMR annexed to the respective LRR).

This result raises the question whether project teams' commitment to evaluability decreases after approval. An analysis of available Loan Contracts (LC) for projects in the sample throws further light on this question. In most cases there is at least a brief and generic commitment to evaluation, such as: "*The Borrower agrees to cooperate in the evaluation of the Program to be carried out by the Bank after the Program's execution, with the purpose of identifying, to the extent possible, whether the objectives of the Program have*

been reached, and to provide to the Bank all the information, data and documentation that the Bank may request to carry out said evaluation."[17] Less than 50% of the LCs analyzed by OVE[18] included a specific mention to the Results Framework. Among them, only two cases did OVE find a specific mention the impact evaluation proposed in the project, which would reinforce the commitment from the borrower's side to collect the specific data proposed in the LD.

C. Analysis of the Criteria in the DEM

As mentioned above, OVE used this first validation exercise as an opportunity to review the DEM criteria and recommend adjustments to Management. The intention in this section is to serve as the basis to provide constructive suggestions to improve the tool in its capacity to measure evaluability. OVE identified six areas which can be strengthened in the SG DEM.

- **First, the DEM has some criteria with greater subjective content which require more guidance.** OVE struggled to validate some of the criteria because of their high level of subjectivity. This was particularly true for the first six questions of the DEM -- the Program Diagnosis sub-section. All the evaluators involved in the exercise received formal training on the DEM and had access to the DEM tutorial. Nevertheless, many times it was impossible to reach consensus on a straight answer, YES or NO, as required by the DEM. OVE evaluators were given the option to point out situations in which neither a straight YES nor a straight NO could be assigned.[19] The intention of giving evaluators such option was to show that in the case of more subjective criteria, the evaluators would often fall into the gray zone. Indeed, in 21.6% of the cases OVE evaluators judged that the criteria could not be easily answered with a clear YES or NO for the projects they were analyzing. Most of these cases (68.75%) involved in the first six questions of the DEM, which constitute the Program Diagnosis Sub-section. In contrast, evaluators found the questions in the Monitoring and Evaluation section to be clearer and more objective.

- **Second, the criteria regarding the indicators are inadequately considered independently of each other.** Currently all questions in the DEM are answered independently, with only a few exceptions[20]. This approach is inappropriate for the criteria involving indicators. The DEM asks whether at least one indicator for each impact/outcome/output has been identified, whether indicators are SMART (specific, measurable, attributable/achievable, relevant and time-bound), and whether they have baselines, targets and means of verification. A project can have indicators that are not SMART, which by definition are not pertinent, and still get credit for having indicator baselines and targets. Similarly, a project can get credit for having SMART indicators even if there is no baseline and thus the indicators will not be useable for evaluation at the end.[21] These criteria should not be considered independently from one another but rather considered as a group.

- **Third, the criteria on internal and external validity are in reality inter-related, and as currently formulated they do not allow for innovation projects to get credit. The DEM has the following two criteria: Criteria #8)** "*Evidence of the effectiveness of the intervention(s) is based on existing evaluations of interventions in other or similar contexts (internal validity)" and criteria #9) "Information about the applicability of the intervention in the country where it is implemented is provided (external validity)." First, they need to be considered together, as external validity is related to the extent to which* internally valid results will be held to be true for other interventions in similar cases. During the validation exercise, OVE found evidence that this is not the case – 3 out of 16 projects did not consider them together, as they did not score positively on internal validity but did so on external validity. In addition, the criteria need to consider how to handle innovative projects. A pioneering project, which defines well the model to be tested and how to do so, may not be less evaluable than a project which is replicating what other interventions have done in the past, yet the current criteria would not give it as high an evaluability score because of the lack of comparable evidence.

- **Fourth, the use of the term impact can lead to confusion, and OVE observed that many projects did not properly identify impacts. The DEM requires identification of impacts and SMART indicators to monitor them. Questions 12 to 17 in the DEM are related to the identification of medium- or long-term impacts and respective indicators. Yet what several projects define as impact of the project is, in fact, an outcome.**[22] This can be in part due to the confusion created by using the term impact, which has been used with "*multiple and sometimes conflicting meanings" (ECG Public Sector GPS 3rd edition).*[23] In addition, impacts and the corresponding indicators have very little weight in the DEM. Questions 12 to 17 add to a total of less than 0.05 of the total score. The expected incentive generated by this is teams giving little importance to this section of the DEM. As a consequence, impacts might be identified, but without much rigor, and as observed above, in many cases it is just the outcome of the project.

- **Fifth, the Risk Management section of the DEM is only a compliance checklist rather than an assessment of the quality of the risk analysis, and the high weight given to this section in the absence of a better assessment mechanism tends to inflate projects' evaluability scores unjustificably.** As described in Review of the DEM in early 2011, the Risk Management section of the DEM "*rates the completeness of the information presented in the project's risk matrix, namely if all risks are identified, mitigation measures have been defined and indicators to monitor mitigation measures have baselines and targets to track their implementation*"[24] (GN-2489-4). This section does not include questions regarding risk assessment quality, although there is a general belief around the Bank that this is the case.[25] In addition, the high weight given to this section tends to inflate projects' evaluability score. Since the Risk Matrix is a mandatory part of project preparation process,[26] most projects tend to score ten – regardless of the quality of the analysis included in the matrix.[27]

- **Sixth, it is difficult to assess a PBL or a PBP using the DEM, since some of the criteria cannot be directly applied.** The Economic Analysis section was particularly problematic in the case of the 5 PBPs in the sample. Although they are required to present some sort of economic analysis, neither a CBA nor a CEA could be properly done, as costs of the proposed reforms could not possibly be measured. In some cases the alternative used by the project teams was to perform the analysis for one component of the project where measurement of costs was possible.[28] Although these are relevant analyses, they cannot be considered sufficient for the project as a whole, as many benefits go beyond these components to which a CBA or CEA could be done. Similarly, the criteria on Results Frameworks were not applied uniformly across the PBPs analyzed. In some cases the conditionalities were accepted as output indicators and in others they were not, and OVE could not find documents related to the standards for accepting or rejecting them as indicators.

Other particularities of PBLs and PBPs are related to how to assess evaluability of a second phase, or of projects that are expected to be complemented by an investment loan. In both cases, there are many issues in applying the DEM directly, as it is unclear whether the DEM should consider the program as a whole or the individual loans separately and independently. In one case, BA-L1021 (SEFB), which aimed at supporting Sustainable Energy Framework for Barbados II, was the second operation of a programmatic series. As such, it was difficult to access its evaluability as a stand-alone project. Some of the diagnosis and logic issues, such as the incomplete discussion on the barriers that prevent the development of the Renewable Energy and Energy Efficiency projects and how they should be addressed, should have been identified from the first operation and required the analysis of both documents for the evaluability assessment. In another case, a PBG, HA-L1065, for the Institutional Transformation and Modernization the Energy Sector, was going to be complemented by an investment grant. Not having the two operations tied together led to difficulties in judging the Results Matrix, which clearly states that "*Policy measures are necessary but by themselves are not sufficient to get the intermediate or end of the period results. They need to be coupled with investments.*"

4

4 Non-Sovereign-Guaranteed Projects

OVE attempted to validate five NSG projects[29] and concluded that the NSG DEM is not an appropriate tool for measuring evaluability. In contrast to the SG criteria in the DEM, the NSG DEM focuses on the "evaluability" of each final outcome indicator – that is, it requires an assessment of each indicator separately and independently.

It does not provide a broad view of project evaluability, as it does not provide an overall diagnosis of the problem to be addressed, an analysis of the context in which the project takes place, or the project's vertical logic. It provides an incomplete analysis of monitoring and evaluation plans and limits teams in the choice of indicators. For these reasons, OVE could not assess project evaluability based on the NSG DEM and, consequently, did not validate the scores assigned to these projects.

This section reviews the adequacy of the NSG DEM with respect to the evaluability standards agreed upon in the Review of the SG and NSG DEM in 2011 (Box 1). This chapter illustrates the issues identified in the NSG tools through careful analysis of the five projects reviewed in this exercise. Most of OVE's main evaluability findings regarding such projects were not reflected in their final evaluability scores presented to the Board.

The analysis presented here takes into consideration the logic of private sector operations and the constraints they face, which affects their evaluability. First, many project teams indicated that information that would be relevant for evaluability is often excluded from project documents because of confidentiality issues. Second, private clients often lack incentives to pay for monitoring and evaluation. If a project provides a learning opportunity, resources from the Bank may be needed to finance an evaluation. Finally, it is difficult to identify final beneficiaries of financial intermediary operations due to the fungibility of money and the size of the loans relative to client portfolios. This leads teams to focus on the financial transaction or on monitoring a client's overall performance rather than the expected outcomes specified in the LD.

A first weakness in NSG project evaluability concerns problem diagnosis. The LD of NSG projects, as described in the Guidelines,[30] are supposed to present the justification for the Bank's intervention based on an analysis of the market context of the project, the market failures to be addressed, and the means to address them. Although NSG operations are not necessarily designed to directly address specific development problems, they are expected to produce externalities with positive impacts on identified development problems. The intervention can be fully justified only if the LD provides a clear explanation of the problem, presenting evidence and describing its causes and dimension. A project that will have benefits only for the borrower can be perfectly evaluable, even if irrelevant to the Bank's development mandate. The NSG loan proposals analyzed by OVE did not provide clear diagnoses of the market failures to be addressed or explain why they exist and how the proposed solution is an appropriate means to address them. (See Boxes 4 and 5 for examples of an SCF project and an OMJ project, respectively.)

Box 4. Access2services Pilot Facility

Project RG-X1132 consists of two different loans: $50 million A-loan (with the possibility of a B-loan) for Banco BIC, in Brazil, and a $50 million A-loan for Banco BICE, in Chile. Since it was developed as a pilot for the Access2services facility, 20% of each loan are expected to be directed towards sub-borrowers in the health and education sectors. The remainder of the funds are meant to be used to finance SMEs in Brazil and green projects in Chile. This projects illustrates many of the evaluability issues observed by OVE in is NSG project analysis.

Diagnosis and Vertical Logic. The evaluability of this project is limited by the poor diagnostic, which hinders its vertical logic. This intervention is based on the premise that access to credit is an important constraint faced by private actors in the targeted sectors, but it does not contain a diagnosis that confirms such hypothesis. The project does not describe the constraints that private actors in these four sectors face, and does not provide evidence that funding is indeed one of them. There is no discussion of the market failure that the project is trying to address, or evidence on whether it in fact exists and is significant. The intervention seems to be justified solely by the view that "there is a significant market opportunity for private sector actors to complement the public health and education services in terms of increasing access to low-and-middle income populations."

Relevance as Pilot. Furthermore, the project is set as a pilot, but does not discuss the hypothesis to be tested with this pilot. It also fails to clearly state how the pilot experience is going to be evaluated and what is expected to be learned before it can be scaled up in the future.

Indicators. Since the loan document does not lay out specific objectives for the project, the DEM Worksheet does not include SMART indicators that will help to fully assess whether this specific intervention is successful or not. For example, while monitoring ROE, ROA and Non-performing Loans is important to ensure that the client remains profitable, solvent and well managed, it provides an incomplete picture of the results of the project itself, giving little insight into the actual developmental outcomes of the intervention. The economic analysis of the project is limited to a calculation of EROE, which is calculated by adjusting ROE to include taxes. No other economic adjustments (such as benefits to consumers, employees, communities, etc) were made in this case.

Although not included in the LD, which from the point of view of evaluability should be a self-contained document, SCF seems to be committed to collecting information on other indicators for health and education projects. For instance, for these projects % of publicly funded patients reached and % of students of low- and middle- income population are expected to be monitored.

M&E. The Monitoring plan contains a list of the quantitative indicators identified in the DEM, but does not outline a plan to collect information and monitor other areas for which a potential impact was described in the DEM (e.g. competition, market expansion, demonstration effects from innovation, among others). The Evaluation plan describes the self-evaluation process through the XPSRs, and states that clients will provide, to the best of their knowledge, information about the outcome indicators contained in the DEM. However, there is no explanation of what sort of information the clients are expected to provide or how they will collect and share them with the Bank.

A second issue is the identification of project objectives. The objectives of most of the reviewed projects were not stated in a specific or consistent manner in the LD. For example, project AR-L1135 claims that "The Credit Line seeks to assist a middle size local financial institution focused on serving mid-size corporate in Argentina." At the same time, it says that "The purpose of the Loan is to expand indirectly the pool of available financing for SMEs in Argentina," Similarly, in RG-X1132, detailed in Box 4, the LD does not present any specific development objectives for the Brazilian SME sector or the Chilean green projects, although these sectors are expected to use most of the resources. It only states that the Access2services pilot facility aims to "direct commercially viable financing through financial institutions (FIs) to high positive social impact sectors." This lack of a specific statement of the objectives of the intervention compromises evaluability, since it is unclear what development outcome is to be evaluated at the end.

Partially as a result of poor diagnoses and lack of specificity in the definition of objectives, the vertical logic of the projects analyzed was not fully developed and justified – a third issue identified in OVE's analysis. These projects did not establish a clear link between the activities that will be performed and the expected development results of the interventions. Although DEM Guidelines state that Loan Proposals

should present a table linking outputs to outcomes, none of the projects reviewed by OVE contained such a table, though it could have greatly improve the transparency of what was being proposed. The documents also failed to provide evidence that the proposed interventions could produce the externalities – i.e. development outcomes – as claimed in the LDs. A clear example is BR-L1313, described in Box 5, which did not provide satisfactory evidence that the intervention would be an effective tool to promote early childhood development among the poor in Brazil.

A fourth issue in project evaluability concerns indicators. Though the DEM worksheet focuses on indicators, OVE found serious problems related to the definition and identification of indicators. Projects contained a good overall assessment of business performance, but lacked proper indicators to measure development impact. Although monitoring of FIs' Return on Equity (ROE), Return on Assets (ROA) and Non-performing loans is essential – and important for the Bank to measure credit risk, the indicators that constitute the Project Business Performance section of the DEM do not help measure the contributions of projects to economic development.

The three SCF operations with Financial Intermediaries (FI) analyzed by OVE provide examples of improper consideration of economic benefits. The analysis was limited to calculating EROE by adjusting ROE for taxes and subsidies, without taking into consideration costs and benefits to all relevant stakeholders. Only results for the FIs were considered, giving no consideration to the broader economic impact of the project. There was very little attempt to measure the changes in the market in which the FIs operated. The OMJ projects made a stronger effort to measure the economic return of interventions through more elaborate models to compute the ERR, though the documents that contained the rational for these calculations were not accessible through the links contained in the DEM Worksheet.

The problems with indicators mentioned above might result in part from the format of the DEM, which limits the number of indicators teams can use. The projects analyzed by OVE included a highly restricted number of indicators. In the first two sections of the DEM -- Business Performance and Development Outcome – the projects included at most 4 indicators. Consultations with project teams revealed that indicators beyond those in the DEM were expected to be monitored. OMJ projects in particular appear to monitor a large number of indicators through PULSE,[31] which seems a promising initiative to improve project evaluability, as it aims to facilitate monitoring and evaluation of projects. OMJ teams interviewed for this evaluation provided the information for their projects included in PULSE, following the indicators defined in IRIS.[32] The tables provided by the teams, not annexed to the projects, are more complete and informative than what is included in the DEM. However, these additional indicators are not included in the LD, which affects the projects' evaluability, as the LD should be self-contained, including the definition of SMART indicators. The DEM appears to create unnecessary constraints that compromise rather than facilitate evaluability.

Box 5. PUPA

Project BR-L1313 consists of a $3 million secured loan from OMJ to launch PUPA, a social business developed by ZOOM Editora – the exclusive representative from LEGO in Brazil – to address gaps in Early Childhood Development (ECD). The PUPA product will consist of a kit including magazines, audio-visual aids and LEGO toys, and the program will include training for caregivers in ECD techniques. The product will be sold to parents and caregivers in low-income communities through a network of partner NGOs and micro-franchisees (self-employed women who will be responsible for signing up customers, helping them select the appropriate PUPA packages, delivering the packages, and conducting surveys to track social performance indicators).

The project is motivated by the evidence that low levels of cognitive development early on in life are associated with lower educational achievement and, consequent, lower wages. The literature on the impacts of early childhood development interventions was also reviewed to justify the support of the Bank in this pioneering approach.

The project is based on the assumption that the proposed intervention, which is an innovation and the product has not yet been developed, will have similar impacts as other ECD interventions. In fact, this initiative consists of a pilot. In the first stage, it will take place in five low-income communities in Sao Paulo in which ZOOM has strong relationships with NGOs that run free daycare facilities. In the second year, PUPA will be launched formally and sold across the state of Sao Paulo. Once operations are well established, the program will be expanded to other regions of the country. This means it represents a good opportunity for learning.

Important evaluability issues appear to exist, based on information presented in the LD and its annexes, First, there is a lack of clarity in the presentation of the model. Many aspects of the intervention seem to be left to be designed after approval. Even the product to be sold has not been developed. Second, this pilot phase does not have a clear evaluation plan. The test to be applied to children to measure their cognitive development is still under construction and it is reliant on a Technical Cooperation (TC) to be approved. So there is a high risk of missing the opportunity to learn from this experience.

Third, the LD lacks a good diagnosis of the market for this product or its likely development impact. It is unclear whether there is real demand for the product and at what the price it can be commercialized among the targeted population - the parents and informal caregivers in the base of the pyramid. In fact, the need for the product is not well articulated in the LD, as the project only provides evidence of low preschool/ daycare attendance, but does not establish a clear link between this and low ECD. The project is based on the hypothesis that being assisted by an informal caregiver results in lower cognitive development, which is an empirical question.

There is a clear need to further strengthen the indicators to assess private sector development in the DEM.

(C) Singal Suhler Moran, 2012

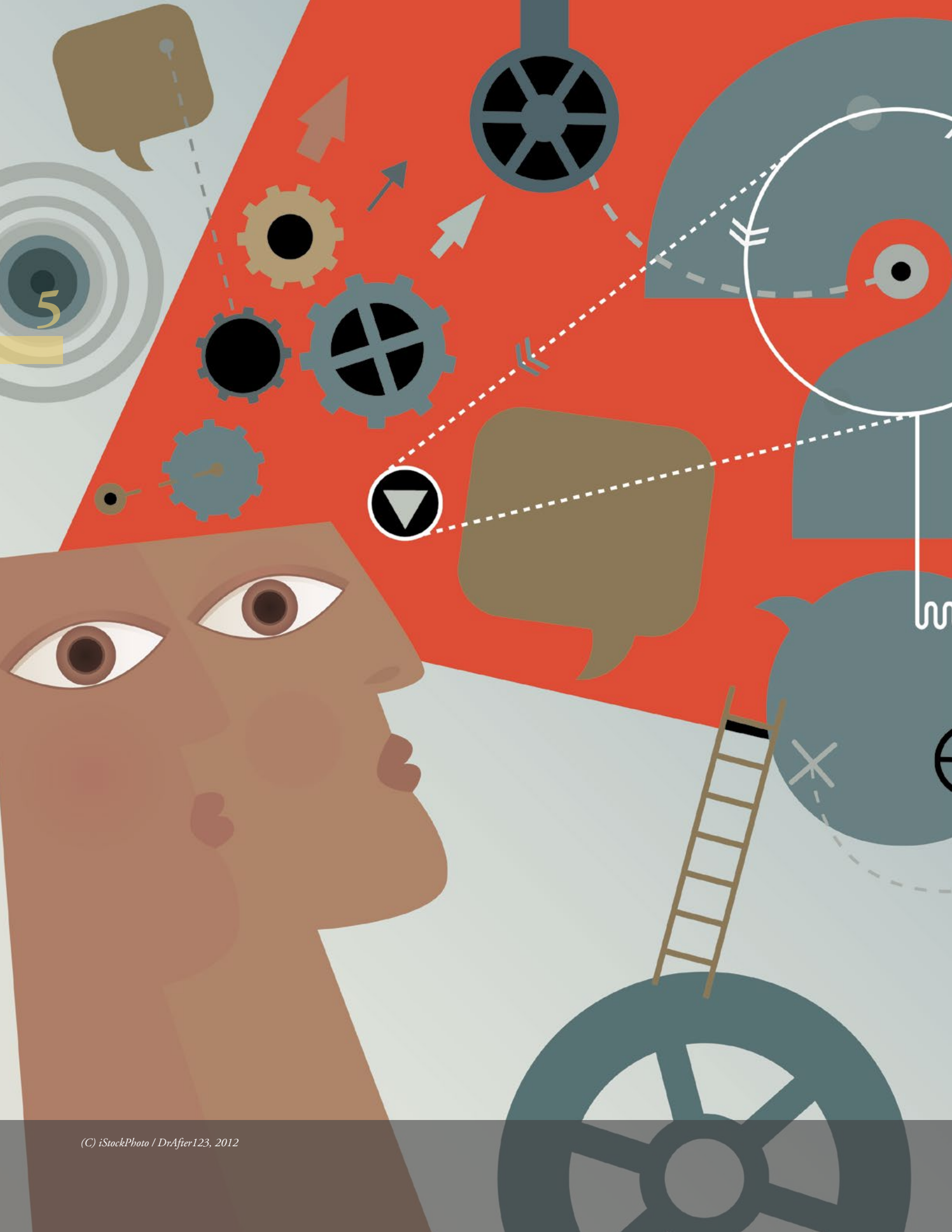
There is a clear need to further strengthen the indicators to assess private sector development in the DEM and to provide more guidance to teams to identify and measure them. Quantitative indicators are required only in the Project Business Performance and Contributions to Economic Development sections, while quantitative indicators and qualitative benchmarks are accepted in the Environmental/Social Risk & Mitigation and Contribution to Private Sector Development sections. The last section of the DEM (Private Sector Development) requires indicators/results for a list of areas: competition, market expansion, private ownership and transition impact, technology and know-how transfer, demonstration effect from innovation, standards for corporate governance, legal and regulatory framework, physical or financial market infrastructure. Yet in the projects analyzed by OVE, most of the sub-sections had no indicators or adequate qualitative benchmarks to assess whether the expected results had been achieved. For example, project AR-L1135 claimed in the market expansion sub-section that *"CMF will have a modest impact on the expansion of the SME market," but did not define an indicator, baseline, or target for this expected expansion, stating merely that "actual expansion will be verified at the time of evaluation […] based on information provided by CMF."*

A fifth issue affecting NSG project evaluability is the DEM scoring system. The application of the NSG DEM in practice generates negative incentives for project teams, as the highest evaluability score is given when a criterion does not apply to the project, inflating the final evaluability score of the project. The categories analyzed in the Private Sector Development Section of the DEM were defined following the

ECG-GPS, which states that "*In assigning ratings for [private sector development], the following factors may be considered:: Competition […]; market expansion […]; private ownership and entrepreneurship […]; frameworks for markets […]; transfer and dispersion of skills […]; demonstration effects […]; standards for corporate governance and business conduct […]; development of financial institutions and financial / capital markets […]; and development of physical infrastructure […]."(ECG-GPS for private sector 3*th edition, underline added for emphasis). *As suggested in the GPS, this is an indicative list of areas that contribute to private sector development. Not all projects are expected to consider all of them. In fact, OVE observed that teams have argued that some of these dimensions do not apply to their project. However, the corresponding evaluability score is ten. During interviews with VPP Development Effectiveness Officers, OVE learned that in the case of non-applicability a score of ten is always assigned, even though this rule is not made explicit in the NSG DEM Guidelines. As a consequence, evaluability scores have been inflated, resulting in almost all projects scoring above 9. Projects are being labeled as highly evaluable despite all the evaluability issues noted above.*

A final issue observed in projects but not addressed in the DEM is the failure to assess the adequacy of monitoring and evaluation. The Guidelines describe the standard monitoring[33] and evaluation[34] procedures for all projects. If the project has evaluation activities that go beyond these standard procedures (e.g. in-depth ex-post economic analyses or impact evaluation), the LD should specify them, along with a description of how the Bank will support the client in monitoring the development results of the project. Yet the DEM does not assess the appropriateness of the evaluation plan or the availability of resources (see Box 5). When the project team does not foresee evaluation activities other than the standard practices (annual supervision reports and XPSRs), the NSG DEM guidelines state that officers are not required to lay out a detailed monitoring and evaluation plan. Only a description of how the indicators included in the project will be monitored through the life of the project is requested. This provision is insufficient. First, the XPSR does not require the provision of evidence of results attributable to the intervention. This means that a complete XPSR is necessary, but not sufficient to show a causal relation between the project and the development impacts. Second, the evidence from OVE's XPSR validation exercises shows that when projects reach early operating maturity, the information necessary for a complete assessment of project performance is often unavailable. Even though none of the projects for which DEMs were prepared has been validated, the analysis of the case studies suggests that the DEM will not be able to resolve the problem highlighted by OVE's XPSR validation team.

Based on the analysis presented above, OVE finds that major revisions are needed in the NSG DEM to improve the quality of the evaluability assessment, and that such changes can also reduce the burden on Investment Officers. The necessary information for assessing project evaluability should be contained in the LD rather than being spread across overlapping documents, which is time consuming for project teams and reduces the clarity of the information assessed by evaluators.



5 Conclusions and Recommendations

The experience of this pilot exercise was fruitful, and OVE intends to continue using validation of DEMs as its main source for reporting on the Bank's project evaluability. In the future, OVE intends to share each validation with Management as it is completed and to report periodically to the Board on a set of validations.

During this first validation exercise, OVE noticed that the tools used to assess evaluability in the Sovereign Guarantee (SG) and Non-Sovereign Guarantee (NSG) projects are very different, and the scores computed using each of them cannot be compared. The SG DEM has a more comprehensive approach to evaluability, which might explain the high variability of scores among the SG projects vis-à-vis the NSG ones. According to the 2011 Development Effectiveness Overview (DEO), all NSG projects were classified as highly evaluable and most of them have evaluability scores concentrated between nine and ten, while among the SG projects evaluability scores are spread between 5.8 and 10.

In validating the DEMs for SG loans, OVE's final scores were similar to those of Management. However, OVE's analysis in this report suggests that the SG DEM can be further improved in a few areas. Most of the differences between OVE's assessment and the DEM scores at approval were concentrated in the first part of the DEM, which contains more subjective criteria. Other sections of the DEM could also be improved to better assess evaluability.

OVE was unable to validate the NSG projects' DEM, as the tool currently used does not properly measure evaluability, and OVE found serious evaluability problems in the projects analyzed. A small number of NSG projects were analyzed in this exercise, but this number was sufficient to reveal basic shortcomings of the instrument. The current NSG DEM does not address these shortcomings, yet the evaluability scores reported for these projects were similar and very high, resulting in an uninformative assessment. The NSG DEM currently fulfills too many roles, and the evaluability

analysis needs to be separated from the assessment of likely project success. At the same time, the indicators to be monitored should be listed together. The complexity of the current NSG DEM format is unnecessary.

Further analysis of the incentives around project design and later monitoring and evaluation suggest that even highly evaluable projects at approval might not be able to report on results at completion. Before projects are approved, the need to prepare the DEM and, more recently, to score five or more in each of the evaluability sections incentivize teams to put more emphasis on evaluation during project design. However, after approval attention can shift towards execution at the expense of finalizing the analysis not concluded in the LD. The evidence in this report suggests that even when the Bank's instruments are effective in orienting projects towards more evaluable interventions, this might not be enough for the Bank to report on results in the future.

RECOMMENDATIONS

OVE makes the following recommendations for Management to further strengthen the DEM tools and their application:

- **Strengthen the DEM for SG operations by:**
  - **Providing more detailed guidance for questions with greater subjective content.** This was particularly true for the first six questions of the DEM, whose tutorial need to be clarified and questions need to be refined.
  - **"Nesting" the questions on indicators rather than answering each question independently**. The DEM asks whether at least one indicator for each impact/outcome/output has been identified, whether indicators are SMART (specific, measurable, attributable/achievable, relevant and time-bound), and whether they have baselines, targets and means of verification. All these questions should be considered in a group.
  - **Reviewing the criteria on external and internal validity.** They need to be considered together, as external validity is related to the extent to which internally valid results will be held to be true for other interventions in similar cases. Also, the tutorial of these criteria needs to consider how to handle innovative projects. A pioneering project, which defines well the model to be tested and how to do so, may not be less evaluable than a project which is replicating what other interventions have done in the past.
  - **Eliminating the requirement to extend the results framework and indicators to ultimate impacts.** Although, from the point of view of vertical logic it is important to identify impacts and relate them to the outcomes, projects should not be required to identify indicators to measure impacts that go beyond the intervention.
  - **Taking the Risk Management Section score out of the overall evaluability assessment and scoring.** Although the risk section is important and should be kept in the DEM for the purpose of assessing project quality, its format is not adequate for measuring evaluability.
  - **Revising the criteria for PBLs and PBPs in three main aspects.**
    - **Rethink the required economic analysis.** Currently, Economic Analysis section is computed as the maximum between the CBA and the CEA. Yet neither a CBA nor a CEA is applicable to PBLs and PBPs. An alternative would be to specify the score of this section as being the maximum score among the CBA, CEA or general economic analysis in the case of PBLs/

PBPs. This general analysis would be expected to provide evidence on the likely economic impacts of the policy and institutional reforms supported by the PBL/PBP.

- **Clarify the standards for output indicators.** In some cases the policy conditions were accepted as output indicators and in others they were not, the standards for accepting or rejecting them as indicators should be made clear.

- **Clarify the unit of analysis.** In particular, the unit of analysis should be well defined in the evaluability standards to clarify whether the DEM should consider the program as a whole or individual tranche of a PBL separately and independently. The same is true for projects that are expected to be complemented by an investment loan.

- **Thoroughly revise the DEM for NSG operations and improve their evaluability by:**

  - **Disconnecting the assessment of evaluability in the DEM from the measurement of likely development impact.** The dual purposes of the current DEM constrain the tool's format and limit the assessment of project evaluability. OVE recommends that IDB develop a more comprehensive evaluability assessment tool, free from the constraints imposed by the current DEM structure.

  - **Ensuring that the revised NSG evaluability tool covers the essential aspects of evaluability.** Evaluability is a straightforward concept that should be applied equally to any kind of intervention. Its assessment has to be grounded in at least three dimensions: project logic, including the diagnosis of the market failure the project is trying to address and clear definition of the objectives the Bank aim to achieve with such intervention; economic analysis, which should attempt to measure all the impacts the intervention will have on the targeted market through a proper analysis of supply and demand; monitoring and evaluation, precisely setting the plan for who is going to be responsible for collecting information and how the information will be used to measure the impact of the intervention.

  - **Including a Results Matrix to clearly present all output and outcome indicators and the links between them.** A Results Matrix would help organize relevant information and show project logic and output and expected outcomes in one place (rather than separately as forced by the current DEM). The NSG DEM Guidelines mandate that projects include an Output Table, which could fulfill this role. However, in practice such table does not appear to be in use.

- **Clearly identify funding for project evaluation prior to approval.** Funding may come from NSG loan proceeds (financed by borrower) or, if this is not feasible, from alternative financing sources, but it needs to be identified up-front.

- **Consider having one unit in IDB oversee evaluability for all projects -- whether SG or NSG -to ensure consistency and comparability of treatment.** One option is that suggested by Management in GN-2489-4: "*In 2012 a review of the effectiveness of this process will be completed and if appropriate SPD will take this responsibility and will also begin to validate the ratings for the NSG DEMs.*" The first recommendation above on DEM content suggests that the SG and NSG DEMs need not be as different as they currently are. Bringing responsibility for both under one roof might help eliminate unnecessary inconsistency and promote learning and innovation that serves both SG and NSG clients.

- **Ensure post-approval follow-up to monitor and assess results downstream.** This report raised an empirical question regarding the incentive structure, which seems to work against a culture of reporting on results. OVE will further address this issue in subsequent evaluations.

Annex

Annex A

Development Effectiveness Matrix for Sovereign Guaranteed Operations

Part II - Evaluability

Fill the white cells of the "Yes/No column" with "Yes" if the Proposal of the project includes the information requested or fulfills the criterion. If the project doesn't fulfill the criterion, leave the cell blank.

In those cells where you are requested to provide referenced information, please provide the references by writing it in the same cell. Fill the "Yes/No column" with "Yes" if you provided the information.

Criterion	DEM TUTORIAL The answer is "yes" if:
Section 3. Program Logic	
Program Diagnosis	
The main problem being addressed by the project is clearly identified	The "Background and Problem Addressed" section of the POD specifies: i) clear, accurate, and adequate information about the situation. ii) the reasons why the situation constitutes a problem requiring a solution. iii) basic quantitative and qualitative data that allow for an adequate dimensioning of the problem.
The intended beneficiary population is clearly identified (households, localities, firms, users, or overall population)	The "Background and Problem Addressed" section of the POD clearly specify the households, localities, firms, persons, among other entities, that are expected to receive good and services delivered by the project, and that fulfill certain characteristics.
The main factors (or causes) contributing to the problem are clearly identified	The "Background and Problem Addressed" section of the POD includes the relationships and interactions that occur among the factors and elements that constitute the problem.
Empirical evidence of the main determinants of the problem is provided (See Guidelines for guidance on what constitute sufficient empirical evidence)	The "Background and Problem Addressed" section of the POD includes the basic quantitative and qualitative data that allow for an adequate dimensioning of the factors.
Magnitudes of deficiencies are provided for main factors (in order to assess the relative importance of identified factors)	The "Background and Problem Addressed" section of the POD includes, in a precise manner, a review of studies or assessments that try to identify and quantify the contribution of the main determinants of the problem. The POD should include the references to these studies.
Diagnosis takes into account specific country characteristics in the area of project intervention	The "Background and Problem Addressed" section of the POD includes, in a precise manner, the historical, political, social, and economic context in which the problem arose, as well as those in which its solution will be attempted.
Proposed Interventions or Solutions	
Proposed Intervention(s) are clearly linked to problems or needs identified in the Diagnosis	There is a clear and precise specification on how the proposed intervention will contribute to solve the problems identified in the diagnosis
Evidence of the effectiveness of the intervention(s) is based on existing evaluations of interventions in other or similar contexts (internal validity)	The "Justification" section of the POD includes empirical evidence (quantitative and qualitative updated data) of the effectiveness of similar interventions. The POD should include the references to these studies.

Development Effectiveness Matrix for Sovereign Guaranteed Operations

Part II - Evaluability

Information about the applicability of the intervention in the country where it is implemented is provided (external validity)	The "Background and Problem Addressed" section of the POD includes, in a precise manner, a review of studies or assessments that addresses the degree to which the intervention would hold in the country in which the intervention will be implemented, and in a certain time and moment. The POD should include the references to these studies.
The dimension of proposed solution is related to the objective of the project and its magnitude	The scope of the interventions (mainly addressed by the results matrix) is clearly related to the magnitude of the problem.
Results Matrix Quality	
Section 1: Vertical Logic	
Verify the vertical logic. Each level logically contributes to the next higher level. Inputs -> Activities -> Outputs -> Outcomes -> Impacts **Top down:** Ask how a particular level can be attained. The answer should be: by successful completion of the immediate lower level **Bottom up:** Ask why a particular level is being done. The answer should be: in order to attain the next higher level.	The results matrix of the project should have a clear definition of the relationships of cause and effect between the different parts of the problem that correspond to the tree levels of objectives: outputs, outcomes and impact. The vertical logic implies the existence of a model in which: The production of outputs will contribute to the achievement of the outcome of the project. From there, the next level implies that if the outcome is achieved, there will be a significant contribution to the achievement of the impact of the project.
Section 2: Impact of the program	
The desired medium- or long-term impacts are stated in the POD and are clearly related to the country strategy/country program results matrix for that sector or area of intervention. In fact, the impact of the project usually is the same as the stated sector objective and indicator for a country strategy result matrix. A single project may generally not the sole means for attaining the general objective; it may only contributes towards the general objective indirectly. Several other country and/ or sector programs are needed to support the achievement of the general objective or desired impact. They are specified as an expected verifiable achievement that is expected as a medium- or long-term result of the intervention.	The results matrix of the project clearly defines the expected impact of the project and if this impact is clearly related to the country strategy objectives for that sector or area of intervention. Impact of the Project: Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended. Impacts generally refer to changes in living conditions or institutions and are not under the direct control of those responsible for project management.
There is at least one indicator is identified for each impact. Indicators are the selected metrics by which it is verified if the desired change is taking place.	The results matrix of the project includes selected metrics by which it is verified if the desired change is taking place.

Development Effectiveness Matrix for Sovereign Guaranteed Operations

Part II - Evaluability

Indicators are SMART (Specific, Measurable, Achievable, Realistic and Time-bound)	The results matrix of the project identifies SMART impact indicator: Specific: Precise and unambiguous. Measurable: The indicator is susceptible of measurement, calculation, or computation, and amenable to independent validation. Achievable/Attributable: The indicator is capable of being attributable to the program. Realistic: The indicator is accurate and related to objectives of the program. Time-bound: A specific time (or several times depending on expected outcomes and the monitoring and evaluation plan) to measure the indicator is set.
Every indicator has a baseline value or a predetermined starting point for subsequent comparison of performance.	The results matrix includes ex ante data of the state of the indicators chosen to monitor and evaluate the impact of the project. Baselines for impact provide ex ante information on conditions that are expected to change as a contribution of the project. Baselines establish where the project is starting from, and are essential for measuring progress and accomplishment at some later date.
Every indicator has a target value. A target is a predetermined level of success that is expected within a specified timeframe. In the case of performance-based loans the accomplishment of targets trigger disbursement, otherwise targets are indicative directions for change.	The results matrix includes a predetermined quantitative level of impact that is expected within a specified timeframe. The target value has to be express in the same baseline unit.
Todos los indicadores tienen una fuente de datos o un plan claro para recopilarlos.	La Matriz de Resultados o el plan de seguimiento y evaluación del proyecto incluyen una fuente definida de datos o un plan claro de recopilación de datos para cada indicador de impacto.
Section 3: Outcomes	
The desired improvements (effects) as a result of the project are clearly stated. The outcome (s) should describe what is expected to be different as a result of the delivery of project outputs have been delivered by the project; NOT what the project is going to do. They are stated as expected, verifiable achievements (i.e. increased reading scores for children, decrease in malnutrition, etc. are project outcomes of cash transfer programs. Potable water access 24/7 can be an outcome of a water project. Lower transportation cost is an outcome of a roads project. Reduced time or cost for legalizing a business is a public sector project outcome).	Outcome: The desired improvements (effects) as a result of the project are clearly stated. The outcome (s) should describe what is expected to be different as a result of the delivery of project outputs; NOT what the project is going to do. They are stated as expected, verifiable achievements (i.e. increased reading scores for children, decrease in malnutrition, etc. are project outcomes of cash transfer programs. Potable water access 24/7 can be an outcome of a water project. Lower transportation cost is an outcome of a roads project. Reduced time or cost for legalizing a business is a public sector project outcome).
There is at least one indicator is identified for each outcome. Indicators are the selected metrics by which it is verified if the desired change is taking place.	The results matrix of the project includes selected metrics by which it is verified if the desired change is taking place.

Development Effectiveness Matrix for Sovereign Guaranteed Operations

Part II - Evaluability

Indicators are SMART (Specific, Measurable, Achievable, Realistic and Time-bound)	The results matrix of the project identifies SMART outcomes indicator.
Every indicator has a baseline value or a predetermined starting point for subsequent comparison of performance.	The results matrix includes ex ante data of the state of the indicators chosen to monitor and evaluate the outcome of the project. Baselines for impact provide ex ante information on conditions that are expected to change as a contribution of the project. Baselines establish where the project is starting from, and are essential for measuring progress and accomplishment at some later date.
Every indicator has a target value. A target is a predetermined level of success that is expected within a specified timeframe. In the case of performance-based loans the accomplishment of targets trigger disbursement, otherwise targets are indicative directions for change.	The results matrix includes a predetermined quantitative level of outcome that is expected within a specified timeframe. The target value has to be express in the same baseline unit.
Every indicator has one source of data, or a clear plan for collecting it.	The results matrix or the monitoring and evaluation plan of the project includes a defined source of data or a clear data collection plan for each outcome indicator.
Section 4: Outputs	
Project deliverables are clearly specified. Outputs are project "deliverables." They summarize what the project is contractually accountable to provide. They are stated as expected, verifiable achievements. (i.e. school access increased, children de-wormed, hectares planted, new procedures operational, personnel trained, # of connections to clean water)	The results matrix clearly defines the goods and services being delivered by the project.
There is at least one indicator is identified for each output. Indicators are the selected metrics by which it is verified if the desired change is taking place.	The results matrix of the project includes selected metrics by which it is verified if the desired change is taking place.
Indicators are SMART (Specific, Measurable, Achievable, Realistic and Time-bound)	The results matrix of the project identifies SMART output indicator: *Specific:* Precise and unambiguous. *Measurable:* The indicator is susceptible of measurement, calculation, or computation, and amenable to independent validation. *Achievable/Attributable:* The indicator is capable of being attributable to the program. *Realistic:* The indicator is accurate and related to objectives of the program. *Time-bound:* A specific time (or several times depending on expected outcomes and the monitoring and evaluation plan) to measure the indicator is set.

Development Effectiveness Matrix for Sovereign Guaranteed Operations

Part II - Evaluability

Every indicator has a baseline value or a predetermined starting point for subsequent comparison of performance.	The results matrix includes ex ante data of the state of the indicators chosen to monitor and evaluate the outputs of the project. Baselines for outputs define the goods and services being delivered prior to the initiation of the project. Baselines establish where the project is starting from, and are essential for measuring progress and accomplishment at some later date. Adequate baselines must specify ex ante the timing, quantity and quality of good and services to be delivered.
Every indicator has a target value. A target is a predetermined level of success that is expected within a specified timeframe. In the case of performance-based loans the accomplishment of targets trigger disbursement, otherwise targets are indicative directions for change.	The results matrix includes a predetermined quantitative level of good and services that is expected to deliver within a specified timeframe. The target value has to be express in the same baseline unit.
Every indicator has one source of data, or a clear plan for collecting it.	The results matrix or the monitoring and evaluation plan of the project includes a defined source of data or a clear data collection plan for each output indicator.
Section 5: Project Monitoring Report (PMR)	
Outputs indicators have annual targets	
Total project costs are grouped by each expected output	
Costs for each output have annual expected amounts	
The sum of the total planned costs for all outputs is equivalent to the total project amount (including counterpart) detailed in the Loan Proposal.	
Section 4. Economic Analysis	
Cost-Benefit Analysis (CBA)	
The project has an ERR and/or NPV for its main components	Information on how the ERR and/or NPV was obtained should be included in the "Economic Analysis Annex" of the POD. This Annex should present the CBA undertaken, indicate what socioeconomic benefits were used in the calculation, present the annual benefit flow for the life of the project, explain (where necessary) the methodology used for benefit quantification, and include the annual flow of project cost (both investment and operation and maintenance) and any adjustment made to reflect real resource costs. The matrix with the ERR and/or NPV calculation must be presented showing the benefit and costs streams used.
The economic benefits are adequately identified and quantified.	See Templates for Economic Analysis
All real resource costs generated by the project during its life are included in the calculation.	See Templates for Economic Analysis
Assumptions used in the analysis are reasonable and clearly spelled out.	See Templates for Economic Analysis
Sensitivity analysis is performed and includes all key variables that could affect project costs, benefits and assumptions.	See Templates for Economic Analysis

Development Effectiveness Matrix for Sovereign Guaranteed Operations

Part II - Evaluability

Cost-Effectiveness (CEA)	
The project has a cost-effectiveness analysis for its main components	Information on how the CEA was undertaken should be included in the "Economic Analysis Annex" of the POD. This Annex should include a comprehensive analysis that explores different alternatives on how a given objective can be achieved while minimizing the present value of the real resource costs involved in achieving this objective. Based on this analysis, a determination should be made as to which alternative is the least-cost means of attaining the desired result.
Key outcomes are adequately identified.	See Templates for Economic Analysis
All available alternatives are considered.	See Templates for Economic Analysis
The economic costs of each alternative are adequately estimated.	See Templates for Economic Analysis
Reasonable assumptions are used in the analysis.	See Templates for Economic Analysis
Sensitivity analysis is performed and includes all key variables that could affect the costs of the alternatives and the assumptions.	See Templates for Economic Analysis
If CBA and CEA are not present	
An acceptable justification is presented for not undertaking a CBA or CEA. Including an adequate discussion of project alternatives and opportunity costs, and of the projects socioeconomic benefits and positive and/or negative externalities	This information needs to be included in the "Economic Analysis Annex" of the POD.
Section 5. Monitoring & Evaluation – Area Rating	
I. Monitoring	
The Bank and borrower have agreed to use the results matrix and the activities defined in the PMR as the Monitoring Plan for the operation	The "Monitoring and Evaluation Plan Annex" of the POD specifies that there is an agreement between the Bank and Borrower to use the results matrix and the activities defined in the Performance Monitoring Report (PMR) as the principal elements for monitoring the operation.
Monitoring mechanisms have been planned and budgeted.	The "Monitoring and Evaluation Plan Annex" of the POD specifies the main activities (PMR and others), timeline (including frequency of data collection) and responsible persons for each activity, and has budgeted each of them though this operation or other financial sources.
Ensure that the source, or means for collecting data (for outcomes, outputs and activities) actually exist, either with the executing agency or/and with the IDB.	The "Monitoring and Evaluation Plan Annex" of the POD specifies the source of data for program monitoring.
II. Evaluation	
General	
The project has an evaluation plan in accordance to the Bank's guidelines for DEM of SG operations	The "Monitoring and Evaluation Plan Annex" of the POD specifies the main activities, timeline (including frequency of data collection) and responsible persons for each activity, and has budgeted each of them though this operation or other financial sources.

Development Effectiveness Matrix for Sovereign Guaranteed Operations

Part II - Evaluability

Timelines are defined to design survey tools/collect baseline/follow up surveys	The "Monitoring and Evaluation Plan Annex" of the POD specifies timelines to design survey tools, collect baseline and follow up surveys.
The evaluation plan has allocated budget	The "Monitoring and Evaluation Plan Annex" of the POD specifies budget for each evaluation activity though this operation or other financial sources.
Methodology to measure incremental benefits Ex post (at completion)	
Method used to evaluate results	
1. Random Assignment	Should be specified in the Monitoring and Evaluation Plan Annex of the POD
2. Non-Experimental Methods (Difference-in-Differences, Propensity Score Matching, Regression Discontinuity, Instrumental Variables, Simulation Model, or other approaches that allow attribution)	Should be specified in the Monitoring and Evaluation Plan Annex of the POD
3. Ex-post Cost-Benefit Analysis	Should be specified in the Monitoring and Evaluation Plan Annex of the POD
4. Ex-post Cost-Effectiveness Analysis	Should be specified in the Monitoring and Evaluation Plan Annex of the POD
5. Before-After or With-Without Comparison (no attribution)	Should be specified in the Monitoring and Evaluation Plan Annex of the POD
Evaluation aspects required to be defined at project design	
A valid comparison/control group has been selected.	A group that is exactly like (statistically equal) the group of participants in all ways except one: their exposure to the program being evaluated. In the case of Before and After method there is a rigorous argument to confirm that in the absence of the intervention, average outcomes would have remained constant over time
The definition of the counterfactual was done at the appropriate level (groups, organizations or individuals), taking into account that it may be necessary to assign groups in order to evaluate (i) interventions with sizeable spillover effects, (ii) interventions delivered to whole groups.	
Power analysis was performed to ensure that meaningful impacts will be detected	The power analysis takes into account the desired Minimum Detectable Effects and the sample size is defined accordingly
The number of waves of data collection and the timing for data collection have been determined	Should be specified in the Monitoring and Evaluation Plan Annex of the POD
The information that needs to be collected (survey questionnaire) and other variables of interest are specified	Should be specified in the Monitoring and Evaluation Plan Annex of the POD

Development Effectiveness Matrix for Sovereign Guaranteed Operations

Part II - Evaluability

Section 6. Risk Management	
Overall risk rate = magnitude of risks*likelihood	Fill in the overall risk rate (PDP validates)
Environmental & social risk classification	Fill in with the environmental & social risk Classification (ESG validates)
Risk Matrix	
Identified risks have been rated for magnitude	
Identified risks have been rated for likelihood	
Mitigation Measures	
Major risks have identified proper mitigation measures	
Mitigation measures have indicators for tracking their implementation	

Annex B

Table I. Projects in the 2011 Sample

Project Number	Project Name	Evaluability Score at Approval
AR-L1124	Strengthening of the Ministry of Economy and Finance	7.20
AR-L1135	CMF Bank - SME Facility	9.37
BA-L1021	Support for Sustainable Energy Framework For Barbados (SEFB) II	8.70
BH-L1028	WSC Support Program - New Providence Water Supply and Sanitation Systems Upgrade	8.60
BO-L1063	Improvement Program for Municipal Management	8.00
BO-L1065	Water and Sanitation Program for Small Localities and Rural Communities	9.60
BR-L1282	Environmental Sanitation Program for Municipios in the Guanabara Bay Area-PSAM	8.30
BR-L1313	PUPA: Innovative Financing for Early Childhood Development	9.67
BR-L1338	BIC Bank - access2services Facility	9.44
CH-L1065	BICE Bank - access2services Facility	9.97
CO-L1103	PBP-Disaster Risk Management and Climate Change Adaptation Program	9.10
DR-L1035	Tourism Development Program - Colonial City of Santo Domingo	9.80
EC-L1087	Electrification Program for rural and marginal urban areas of Ecuador	8.00
EC-L1098	Citizen Security Program: Strength. Pol. Effect. through Improvement in Management & Crime information	8.80
EC-L1099	National Urban Development Program	8.10
ES-L1053	Financing for Access to Drinking Water and Sanitation Systems (FIHIDROS)	8.27
HA-L1065	Institutional Transformation and Modernization Program of the Energy Sector I	6.30
HO-L1079	Support Program for Reform of Pension Institutions and the HHRR Management System	8.60
PN-L1070	Program to Reduce Vulnerability to Natural Disasters and Climate Change I	9.10
PR-L1061	Cadastre and Property Registry Program II	10.00
PR-L1066	Program to Support Job Placement	9.40

Endnotes

1 Leviton et al. (2010).

2 Details about the evaluability tools developed by OVE in 2001 can be found in the Working Paper OVE WP-01/00 of December 2000.

3 The results of the three moments in which OVE assessed project's evaluability can be found in RE-275, RE-333 and RE-379.

4 The NSG DEM was defined in the SCF Development Effectiveness Framework (GN-2473-1) and is based on ECG-GPS. It has been adopted by SCF and OMJ, while the SG DEM was developed by SPD and its original guidelines can be found in GN-2489.

5 In this new process SPD, in the case of SG projects, or the VPP Development Effectiveness Officers, in the case of NSG projects, are expected to provide comments to the teams at the ERM and the QRR. Before projects are sent to OPC SPD or the VPP Development Officers validate the final version of the DEM and prepare an evaluability note to be annexed to the Project Proposal.

6 GN-2489-4 details how each dimension of evaluability should be analyzed.

7 In this report there has been no attempt to compare the findings with OVE's previous evaluability assessments (RE-275, RE-333 and RE-379) because a different tool is being used in this exercise and the projects' evaluability scores, therefore, are not comparable.

8 GN-2489-6.

9 The 2011 revision of the DEM resulted in many of the criteria being revised and broken down into more fine questions, which is more suitable to a checklist. Since then, project teams are expected to provide references to all substantive questions of the DEM. In interview with staff members, it has been identified that the structure of the DEM helped guiding teams to write their LPs, which could help projects to be more evaluable.

10 Even though the NSG DEM Guidelines state that the Loan Proposal and other sections of the DEM should contain information that is essential for the evaluability of the project, these parts of the loan proposal package are not taken into account in the overall score, as they include much more information than the DEM.

11 The basis of the evaluability assessments was the Evaluability Standards included in last revision of the DEM (GN-2489-5).

12 For instance, in the SG operations case, OVE evaluators would validate the criteria, following the tutorial provided by SPD, but would signal in their comments when a straight YES (or NO) could not be assigned. In these cases, they were given the option of rating the criteria with a YES- (or NO+). The incidence of cases with ratings falling in this gray category is reported in Chapter 3. In the case of NSG loans, a more comprehensive analysis of the DEM had to be performed, as detailed in Chapter 4, and evaluators were responsible to point out evaluability issues not considered in the NSG DEM.

13 Among PBLs and PBPs approved in 2011, 20 out of 24 were approved in the second half of the year.

[14] The main difference between OVE's score and the scores validated by SPD in the case of BH-L1028 is related to the Project Diagnosis. The LD identifies four main problems, but at least 2 of them are not well defined. With regard to the first problem identified, related to reduction of water losses, is not clear whether the losses are caused by technical or commercial (illegal connections) reasons, for example. For the problem related to sewerage infrastructure, the second problem identified, the document lacks a discussion about the coverage and performance of the system, what needs to be renewed, and why - since the system is not old (the project states that it is only 20 years old). OVE also identified problems in the analysis of the root-causes and, consequently, the vertical logic, which are related to the poor identification of the problems the project intends to address. In the case of EC-L1099, most of the differences between OVE's assessment and the scores validated by SPD are related to the Proposed Intervention and Quality of the Results Matrix. In particular, it is expected that housing subsidies will improve the living conditions of the families, and that the fact that the families are living in better houses will lead to an improvement in sanitary conditions. However, the relationship between the subsidies and the improvement in living conditions depends on the new houses being built by private sector agents, which is outside the scope of the program. In addition, the Results Matrix does not explain the link between better housing and income increase in a satisfactory way. The project mentions better access to the jobs market and savings in health and services, but the relationship between the impact and results is not clear.

[15] OVE received comments from SPD on the validation forms for these two projects, the ratings included in the table took in consideration the comments and the revised form is available upon request.

[16] Among the 46 LRRs found by OVE, 18 were related to projects that did not need to comply with the NPC because they entered the pipeline prior to October 1st, 2007. Other 3 LRRs correspond to projects which had disbursed more than 60% by the time that the LRR should have been written and following the rules of the NPC would not require and LRR.

[17] This paragraph was taken from the LC for BA-L1021. Similar statements could be found in the LCs for BO-L1063, DR-L1025, EC-L1099, HO-L1079, PN-L1070, PR-L1061, PR-L1066. Other LCs generally mention that the Results Framework or the Policy Matrix should be followed, but do not provide much more information about how the data will be used in the evaluation.

[18] Only 13 Loan Contracts had been signed at the time of this analysis.

[19] The two intermediate categories were given as an option to these evaluators, Yes- and No+, which were supposed to be used when the evaluators judged that either most or some of the information was in the project, but not enough for the criteria to be rated as a YES.

[20] When answering the DEM teams should either feel in the CBA section or CEA, the score of the Economic Analysis section is computed as the maximum score between the two. The same is done in the case of the method chosen by teams for the evaluation. More than one method might be chosen, and the score is given as the maximum among the options chosen: the magnitude of the score goes up from "before and after without attribution" to "random assignment."

[21] In analyzing the 16 projects OVE found some examples of non-smart indicators, which received credit for having baseline/target/ and or means of verification. This is the case of AR-L1124, BA-L1021, BO-L1063, EC-L1087 and EC-L1098. In BR-L1282, all indicators were identified as SMART, but did not means of verification, which might invalidate them at the end.

[22] This problem was identified in almost half of the projects. For instance, BR-L1282, the Environmental Sanitation Program for Municipios in the Guanabara Bay Area-PSAM, does not properly discuss the effects of less pollution in the Guanabara Bay, and defines as an impact indicator the decrease of wastewater disposed in the bay, which is an outcome of this intervention. Another example can be found in PR-L1066, the Program for Job Placement in Paraguay, the project's expected impact of "contributing to improved job placement" is aligned with the CS, but it is too generic and not measured anywhere. The provided indicators in the Results Matrix are outcome indicators.

[23] The DEM tutorial has the following definition: "Impact of the Project: Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended. Impacts generally refer to changes in living conditions or institutions and are not under the direct control of those responsible for project management." The term impact evaluation is usually based on attribution and measures the changes in outcomes. For instance the International Initiative for Impact Evaluation (3ie) defines it in the following way: "Impact evaluations measure the net change in outcomes amongst a particular group, or groups, of people that can be attributed to a specific program using the best methodology available, feasible and appropriate to the evaluation question(s) being investigated and to the specific context." Outcome is defined in the latest public sector ECG GPS as the final level in the results chain, reflecting the objectives of the project, which is similar to the definition in the DEM "The outcome (s) should describe what is expected to be different as a result of the delivery of project outputs."

[24] This section is currently composed by four criteria: i) identified risks have been rated for magnitude; ii) identified risks have been rated for likelihood; iii) major risks have proper mitigation measures; iv) mitigation measures have indicators for tracking their implementation. Except for the third criterion, there is no assessment of the quality of the risk analysis included in the project.

[25] For instance, in the message sent by EVP to Bank staff members on April 28, 2011, regarding Risk Procedures and Fiduciary Arrangements it was stated that "The DEM will also continue to review the quality of the risk assessment." This confusion might be a result of the fact that when revising this section of the DEM the intention was to "provide the Board of Executive Directors with an overall assessment of the severity of the risks identified for the project."

[26] The Procedures for Processing Sovereign Guaranteed Operations updated in July 2010, lists all the necessary parts of the POD to be sent for QRR, among them it is the Updated Risk Assessment Template, "which is not part of the document that will be approved, but it is an essential tool to facilitate the quality of risk review process."

[27] Among the 122 projects approved in 2011 and analyzed in the DEO, almost 70% scored ten, while two projects scored five and the remainder scored 7.5, mostly due to the lack of indicators to track the mitigation of the identified risks. Given that each section has equal weight, scoring ten in the Risk Management section implies already reaching 2.5 of the total score, which is 50% of the threshold. Although GN-2489-4 argues that "the evaluability score is robust to the specific weights," giving ten in the Risk Management section to most projects without an assessment of quality inflates the score of those projects with poor risk analysis. While the 2011 DEO defined that about 90% of the projects were classified as highly evaluable, if the same grouping were done for projects when considering the evaluability score as only the average of the first three sections (i.e. excluding the risk management section) only approximately 80% would fall in the same category.

[28] CO-L1103 is a clear example. In this project a satisfactory CBA was performed for the third component, which is related to the investment to be done by the Fondo de Adaptación (Adaptation Fund). Although it is a relevant analysis, it is not an assessment of the project as a whole, and, therefore, it is incomplete.

[29] The instructions presented in the NSG DEM Guidelines are to some extent imprecise, giving a superficial explanation of how to rate evaluability. There is a lack of specific instructions about what each category entails, what constitutes an appropriate analysis, what kind of evidence can be considered sufficient, among other issues. For example, for the Environmental/Social Risk and Private Sector Development sections, the guidelines instruct investment officers to include a narrative summary of the expected development outcome, as well as diagnostics and analysis to support these expectations and the DEM ratings, but do not explain the level of detail that is expected. It is not clear whether it should be simply a statement of the expected outcome or if it should also describe how the project will lead to the results. In addition, besides quantitative indicators for these outcomes, the guidelines also accept "qualitative benchmarks," but do not give examples of benchmarks that would be considered appropriate. There is also no reference of what can be considered an adequate monitoring and evaluation plan for each of the qualitative indicators.

[30] Guidelines for SCF Staff (Draft) - Using the Development Effectiveness Matrix ("DEM"), version 5.0 February, 2011.

[31] To facilitate the monitoring and evaluation of projects, OMJ started last year to use its own tracking system, PULSE. All indicators agreed between OMJ and clients are included in PULSE, which can be easily accessed and updated by clients. According to the description in the DEO, "clients are able to report directly and quickly on the results of their projects, allowing OMJ and PMU officers to monitor closely social and financial performance metrics." The system includes information on the indicators, their definition and reporting frequency. Over time it builds the history of projects, with data added yearly, semi-annually or even quarterly.

[32] Another effort to improve the efficiency of monitoring and reporting on results was the adoption of the Impact Reporting and Investment Standards (IRIS) taxonomy, which happened in parallel to the implementation of PULSE. IRIS is "a common language for describing the social and environmental performance of an organization." It has been adopted already by many organizations and help making interventions more easily comparable. However, there has been no analysis of whether such indicators are indeed SMART indicators. Although a common language is attractive, Investment Officers should only adopt indicators that are relevant, which can contribute to the understanding of how the project help achieving the proposed results.

[33] Monitoring activities: annual project supervision exercises based on the DEM, with preparation of Project Supervision Reports

[34] Evaluation activities: self-evaluation exercises through the preparation of the Expanded Project Supervision Reports (XPSRs) when projects reach early operating maturity. XPSRs are independently validated by OVE.