RE-308

Ex Post Project Evaluations 2004 Annual Report

Office of Evaluation and Oversight, OVE

Inter-American Development Bank
Washington, D.C.
August 10, 2005

TABLE OF CONTENTS

ANNEXES Available at: http://idbdocs.iadb.org/WSdocs/GEtdocument.aspx?docnum=579515

INTRODUCTION

The Bank's Board approved a new Ex Post Policy (EPP) in October 2003, which mandated two new tasks to OVE: the review and validation of Project Completion Reports and the implementation of ex post project evaluations.[1] This report's scope is limited to the ex post project evaluation task.[2] It's overall purpose is to present the preliminary findings from the first year's experience in implementing EPP. The preliminary findings draw upon three themes: (i) the individual projects evaluated; (ii) the systematic features of the Bank's monitoring and evaluation system; and (iii) evaluative method standards.

It is important to keep in mind that EPP itself was a response to the perceived development effectiveness evaluation gap that had been created when the Bank's policy of obligatory ex post evaluations was abandoned in 1994, and the Operations Evaluations Office closed.[3] The perceived gap has been further brought to the fore by a number of recent initiatives (see Box1) by the IDB to improve the development effectiveness of its operations that together reflect its commitment to a "managing for results" business model. These recent set of efforts are part of the IDB's response to the Bank's Governors' 2002 request for "*all Bank operations to have clear targets that make it possible to define benchmarks to assess our institution's effectiveness*" and a "*move away from an approval- based culture [towards…] a results-based approach*" [4]

Box 1: Recent Initiatives To Improve Development Effectiveness

These initiatives included:

- Approved a new ex post evaluation policy
- Set up a system aimed at improving the quality of entry of projects, plus defining a road map for the development effectiveness agenda, enshrined in a Medium Term Action Plan, MTAP.
- Revamped two of the Bank's main monitoring and evaluation reports[1] with a view of enhancing development outcome information, and retrofitting of a number of projects' information that were found wanting, and improving results reporting in the Bank's Annual Portfolio Management Performance Report,
- Adopted a new disclosure policy that provides for the full public discloser of all evaluation documents sent to the Bank's Board of Directors,
- Placed in each Regional Department's front office a staff member whose primary responsibility is to ensure that the initiatives become a reality on the ground, and created an Office of Development Effectiveness that was later consolidated into a new department of Development Effectiveness and Strategic Planning (DEV).

For a description of these initiatives see "Medium-term Action Plan for Development Effectiveness at the IDB", GN-2324, 2nd of August 2004, "Progress Report on the Management's Actions and future Actions to enhance the Bank's Development Effectiveness", CA-456, 28th of March, 2004, and the Report of the Chairman of the Policy and Evaluation Committee" and " Program to Implement the External Pillar of the MTAP for Development Effectiveness (PRODEV)" GN-2346-2, March 2005. For the extension of the loan facility, PRODEF, to include financing of ex post evaluations see: GN-2351, December 2004.

[1] For the new policy see Ex Post Evaluation of Operations. Final Version, GN-2254-5, and Background Policy Document Ex-Post Evaluation of Operations, GN-2254-6, September 2003. Report of the Chair, Policy and Evaluation of the Board of Executive Directors, GN-2254-7, 8th of October 2003, and OP-305.

[2] Note that amongst the projects evaluated are the "pilot studies" that were proposed in RE-257, September 2003.

[3] Prior to 1994 all of IDB projects had to be evaluated by the borrower ("Modifications to the Borrower's Post Evaluation System, March 1993, CON/OEO-84/93). The reform changed the obligatory system in favour of a voluntary one. Simultaneously, the Operations Evaluations Office, charged with self-evaluations, was closed down.

[4] "Remarks by Mr. Enrique V. Iglesias at the Closing Session" Annual Meeting of the Board of Governors, Fortaleza, AB-2281, 26th of March 2003.

These initiatives also reflect the Bank's commitment to greater accountability as embodied in the joint statement "Measuring, Monitoring and Managing for Development Results" of 2002 and imply an increased demand for evaluation to meet the challenges of a new external (greater accountability demands from stakeholders) and internal (shift towards results based management) context. Managing for results implies the accumulation of knowledge that can be drawn upon to support evidence-based decision-making regarding policy, strategy and project design.

The Report is structured as follows. Chapter I presents a summary of the design features of EPP. Chapter II contains a narrative of the actions taken by OVE to put into practice EPP, including a discussion of evaluative standards. In Chapter III, is a summary of the key findings of a selection of the projects evaluated in the 2004"s ex post evaluation cycle.[5] Chapter IV draws on the project evaluation case studies plus additional research on the systematic features of the Bank's monitoring and evaluation system and presents lessons learned during the 2004 cycle. The final Chapter presents the conclusions and recommendations.

[5] EPP also mentions output of the Borrowers Ex Post Evaluation system, although the report of this exercise is expected after the project closes; presumably data is being collected for it during project execution.

I. EX POST EVALUATION POLICY

1.1 To place the Report's main findings in context, it is useful to set out the parameters that the EPP (GN-2254-5) and other Board approved guidance (RE-284) set regarding ex post evaluation in 2004. The information that was available in 2003 and that went into the design is presented in the policies background paper.

1.2 The EPP endeavours to assign responsibility for evaluation to different actors. Management is assigned the task of developing and maintaining output and outcome indicators during execution, while OVE is assigned to post-execution evaluation (Para. 1.2). Further, "*Borrowers have the main responsibility of collecting the basic information required for the preparation of the PCR and ex post evaluation. Such information is systematically gathered for all projects up to the level of outcomes."(Para.1.6). "…As an integral part of the execution of all projects, the executor must collect the information and indicators for monitoring and evaluation agreed during project design (Para 1.6). OVE … will perform, on a sample basis, in-depth ex-post evaluation of results (impact and/or outcome) of IDB-financed operations, two or more years after completion*." (Para. 1.3) In addition, the EPP states that during project execution "*monitoring and evaluation activities at the IDB are a shared responsibility between the Bank and the country.*"(Para.1.3)

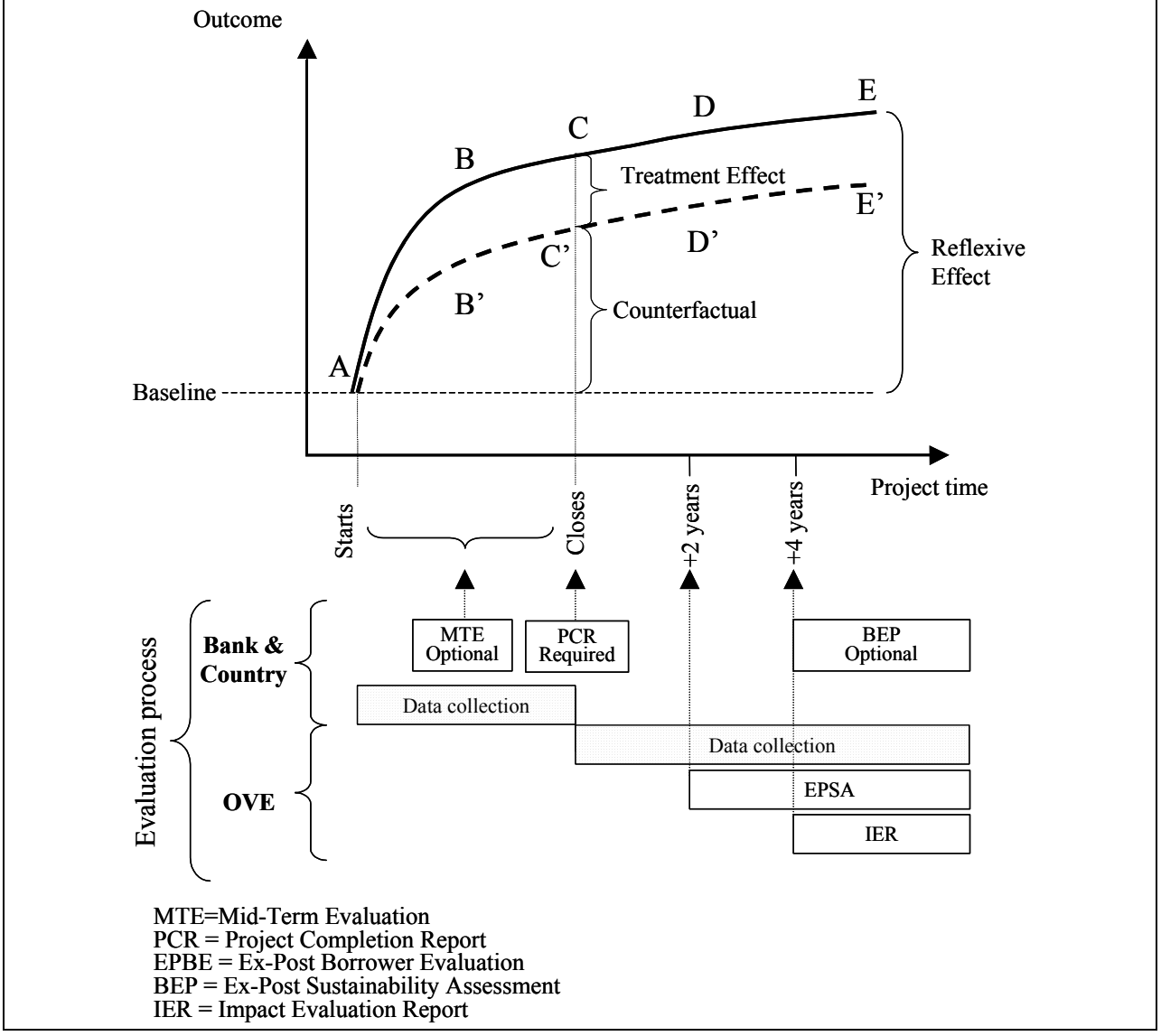
1.3 The policy mentions five reports (see paragraphs 4.1 to 4.5) expected from the system and described in Box 1.1: (i) Project Completion Report (PCR); (ii) PCR Review; (iii) Ex post Performance and Sustainability Assessments (EPSA); (iv) Impact Evaluation Reports (IER); and (v) Borrowers Ex Post Evaluation.

Box 1.1: Reports Required by the EPP

Type of Report	Number	Definition and Evaluative Questions	Responsibility
Project Completion Reports (PCR)	100% of completed operations	Summarize execution experience, information on outputs and outcomes, performance ratings	Management
PCR Review	100% of completed operations	Validation of PCR	OVE
Ex Post Performance and Sustainability Assessments (EPSA)	20% of closed projects per year, 2 or more years after closure	Assess short to medium term results of a program, including outputs and outcomes, institutional sustainability and flow of outputs.	OVE
Impact Evaluation Reports (IER)	At least 2 per year, 4 or more years after closure	Assess long-term results, including impact and outcomes, and effectiveness, efficiency and benefit flow of outcomes.	OVE
Borrower's Ex Post Evaluation (BEP)	Voluntary	Assess results and sustainability.	Borrowing country

1.4 The system described by EPP can be schematically represented as in Chart 1.1. Projects at the approval stage are designed with clear indicators of their development outcome objectives (the "outcome" of the vertical axis). Near the time of project start up, baseline information on those outcomes are collected, i.e. the status of the development outcome prior to the project (value "A"). During project execution further information on the development indicators may be gathered, B for a mid-term evaluation MTE. Data on the outcome is collected near the end of the project (value "C"). The information A, B, C are inputs for statements made in Project Completion Reports regarding development outcome effects, where all outcome changes are reported and perhaps interpreted as automatically attributable to the intervention.

Chart 1.1: Schematic Presentation of the Monitoring and Evaluation System Conceived in EPP



MTE=Mid-Term Evaluation
PCR = Project Completion Report
EPBE = Ex-Post Borrower Evaluation
BEP = Ex-Post Sustainability Assessment
IER = Impact Evaluation Report

1.5 The EPP asks OVE to undertake "impact" evaluations after the project has concluded. Assessing impact means taking a careful look at outcomes in order to separate those

caused by the treatment (the operation) from those caused by other forces operating on the same situations over the same period of time. Assessing these treatment effects requires the collection of additional data during project execution on non-beneficiaries so that values B' and C' can be estimated. Armed with this information (A, B, B', C, and C'), OVE's task is to collect information at point "D" (and estimate D') for an EPSA and data at point E (and E') for an IER.[1]

1.6 Under this interpretation, the main distinction between EPSA and IER is timing. The EPP indicates that an EPSA is conducted at least two years after project completion and an IER at least four years after completion. While this distinction may be based on an intention to measure short - and long term effects respectively, in line with DAC evaluative principles,[2] the heterogeneity of the Banks' projects makes it unlikely that all projects will have short- and long-term over the same time frame. Further, the relevance of the clause for lesson learning and adaptation of policies and programs to improve development effectiveness may require further examination. For example, the logic of the model suggests that a treatment effect could be visible during project execution, particularly for projects that roll out similar treatments for different populations as the project matures. In such cases, there might be learning value in assessing treatment impacts during execution as well as after final disbursement

1.7 The frequent use in the policy document of the conditional in its language suggests that the policy could also be read as a prospective and normative statement: not what is but what should be. Under this interpretation, the EPP implicitly recognises that although the system may not hold for the existing stock of closed projects it should hold for future projects. With this reading, the statement "*... OVE will attempt to identify operations that will be subject to this type of evaluations* (impact) *as early as possible during the project cycle (i.e. during programming and design of the operation).*"(Para. 3.2) makes sense. This implies the creation of a coordination mechanism between the Administration and OVE that ensures that the information necessary for ex post evaluation is collected concurrent to the life of the project. However, the EPP still mandates the evaluation of the existing stock of IDB financed projects.

1.8 The EPP thus defines a system in which all actors must perform as expected. If one actor does not carry out the role assigned, it has negative ramifications downstream. If a project does not have a development outcome(s) metric(s) defined at the approval stage and/or if the evaluative information promised during project execution is not effectively generated, collected and maintained by the Bank for third party use, the ex post exercise as conceived by EPP will be difficult or impossible to do. This is the case for projects approved and closed prior to the approval of the EPP. Under these circumstances the entire evaluative exercise is shifted to OVE. OVE has to define the outcome metrics and to search for information on them back in time: from C to B to A and estimate D and E.

[1] EPP also mentions output of the Borrowers Ex Post Evaluation system, although the report of this exercise is expected after the project closes; presumably data is being collected for it during project execution.
[2] See: "Principles for Evaluation of Development Assistance"; "Review of DAC Principles For Evaluation of Development Assistance", OECD, 1998; and Glossary of Key Terms in Evaluation and Results Management" inhttp://www.oecd.org/document/

Given that the average execution period of a project is about ten years, the evaluative task becomes more akin to an archaeological endeavour: searching and digging for information.

1.9 The EPP left the details of how to implement the ex post task to its practice. The policy's background document (GN-2254-6) offers, *"... the Bank (management and OVE) will prepare operational and methodological guidelines to ensure that homogenous standards are applied throughout the process. The methodological guidelines should include a clear presentation of the different methodologies available, their information requirements, the institutional capacity required for their successful application and an estimate of their cost range. An important definition will be the minimum evaluation standard to be considered acceptable as an ex post evaluation for the Bank..."* (Paragraph 4.8). This report provides preliminary inputs for such guidelines.[3] Methods in general and the minimum method standard in particular are discussed in Chapter II (their data requirements are detailed in Annex 2). Information on costs is given in Chapter III.

1.10 The EPP and its background document were approved concurrently with a plan to implement six pilot ex-post evaluations (RE-284, RE-284-1, RE-284-2) with the objective of "providing guidance in key operative questions that will affect the successful implementation of the Bank's new EPP."[4] The key questions proposed in RE-284 included: (i) how best to coordinate the execution of the ex-post evaluations in order to maximize value added? (ii) what are the budget and data implications of different methodological approaches to ex-post evaluation?; (iii) what are the relative contributions of reflexive and impact evaluations with respect to determining project effectiveness?; and (iv) how should lessons learned best be disseminated within the Bank and borrowing countries? Among its comments to the approved proposal (RE-284-1), Management noted the difficulties in analytically separating EPSA from IER and added that attention should be given the evaluative questions posed in the EPP during the pilot. The six ex post pilot evaluations were co-financed by OVE, RE1, RE3 and DPP.

1.11 The six pilot projects approved in October 2003 were incorporated into OVE's 2004 work program, along with a work plan for the conduct of ten additional ex post evaluations (defined as IER or EPSA) as mandated in the EPP. Given that the evaluative, operative and methodological questions to be addressed by both sets of ex post evaluations are similar and that 2004 was the first year of the implementation of the EPP, all 16 evaluations can be considered part of the pilot and the findings presented in this report refer to the entire set of ex post evaluations conducted in 2004, which allows for a sounder platform for the analysis and recommendations presented. Chapter II details the actions taken to implement ex post evaluations during 2004.

---

[3] "Preliminary" because OVE considers one more year's exercise will generate a more solid information base for such guidelines.

[4] In their comments in RE-284-2, the Policy and Evaluation Committee of the Board agreed that the pilot program "would be used to fine-tune the methodology used by OVE for conducting future impact evaluations." Further, the Committee "looked forward to hearing OVE's recommendations" on "how to ensure the Bank got good value for the resources to be used in such evaluations, how to ensure that project level lessons were effectively disseminated, how best to use these exercises to coordinate with and strengthen borrowers' own evaluation procedures."

II. ACTIONS TAKEN TO IMPLEMENT THE EPP

2.1 To meet the new mandate, OVE took a number of steps to implement the EPP for the 2004 cycle and this chapter documents those activities. Such activities need to be viewed within the overall task of setting up a system for doing a large number of project evaluations each year. Constructing such a system remains a work-in-progress.

2.2 The actions taken in 2004 included: (i) determining the scope of the evaluative questions; (ii) defining a hierarchy of evaluative methods; (iii) defining the sample selection criteria of projects to be evaluated; and (iv) conducting ex post evaluations on 16 projects. The chapter identifies those projects selected for the 2004 evaluation cycle, and contains a description of the implementation mechanisms adopted in the 2004 cycle as well as some measures taken in preparation for the 2005 exercise.

2.3 **Evaluative questions.** The general evaluative questions proposed by EPP are first "…*the extent to which the development objectives of IDB-financed projects have been attained."* and second *"... the efficiency with which those objectives have been attained*" (para1.1)[5]. The evaluative questions' scope should be defined in a way that contributes value added to the information already generated by the Bank. As originally designed, this system remains one that generates information regarding implementation, operational and process aspects of its operations, i.e. the information required to improve the efficiency of implementation of a project and data for accountability to stakeholders. Recently, a requirement to monitor outcome data has been set, with the assumption that this is sufficient to assert something regarding the development effectiveness of the intervention. It may not be sufficient for such a task. Although monitoring of outcome data is important for program managers and generates information on the evolution of a given indicator over time, it lacks the information needed to establish the impact of the project on that outcome or provide explanations regarding the "why" of the changes in the outcome indicator. This is the gap that in-depth, treatment or impact evaluation fills and is the focus of the ex post evaluations undertaken by OVE in 2004.

2.4 **Method.** The methodology employed was chosen so as to maximise the additional information that OVE could bring to the table. In order to answer the evaluative questions posed, a method must be selected in which quantitative statements regarding the evolution of development outcomes (both those explicitly stated as expected as well as unintended) that are attributable to the project can be made with a degree of confidence. Given the key operative questions posed in RE-284, OVE focused its ex post evaluations on two methods: (i) the reflexive method, which assumes that all the changes in development outcomes are attributable to the intervention implied by the Bank's project; and (ii) the treatment effect which attempts to separate out from the gross effects those

[5] Cost efficiency analyses were not included in the evaluations in the 2004 cycle but will be for the 2005 cycle. An attempt at cost benefit analysis has been made for two projects in preparation for the 2005 cycle. In both cases, there was an ex ante exercise, and in one case both ex ante and an ex post evaluations were available. They reveal that without the availability of a methodological reference and follow-up data collection, an ex post evaluation of similar characteristics will encounter difficulties.

that are due to the intervention by establishing causality between the intervention and outcomes in quantitative terms.[6] Here it is useful to inspect again Figure 1.1. The magnitude of the reflexive effect is pictured as the distance between the baseline and the solid line. The magnitude of the treatment effect is depicted as the distance between the dashed line and the solid line. The difference between the two magnitudes—the distance between the baseline and the dashed line—measures non-program factors that also had incidence on the outcome. This is known as the counterfactual, and represents what would have happened in the absence of the program. It is also useful, for clarity, to mention that in the evaluation literature different terms are used to denote the reflexive and treatment effects, and their corresponding evaluations. In particular, evaluations that do not attempt to parse out treatment and counterfactuals are referred to in the literature as reflexive or naïve evaluations and the effects that they measure are known as naive, reflexive or gross effects. Evaluations that measure treatment effects are known as impact evaluations; treatment effects are also commonly referred to as net effects. In any case, to the extent that a project's reflexive and treatment effect estimate differ, only the latter should be used to calculate cost effectiveness of the intervention. In the 2004 cycle, the treatment effect method was taken as the minimum standard, but with the additional task to compare and contrast the gross (naïve, reflexive) and net effects (treatment, impact) estimates.

2.5 The objective of this focus was to illustrate the differences between reflexive and treatment effect calculations, in response to Management's query on the topic made in RE-284-1.[7] A treatment effect evaluation (both concurrent to the life of a project as well as ex post) differs with respect to reflexive evaluation in three key aspects: their primary goal (evaluative questions that can be answered using the method), what and how it is analysed; and the intensity and scope of data collection (see Box 2.1). They are not mutually exclusive or substitutable; both are required to obtain a full picture of a project's results.

Box 2.1: Performance Monitoring versus Treatment Effect Evaluation	
Reflexive evaluation	**Treatment effect evaluation**
Primary goal: accountability to stakeholders and resolution of execution problems, cost-efficiency.	Primary goal: knowledge creation (understanding and improving program treatment effects), cost-effectiveness and cost-benefit.
Analysis of outputs & gross outcome effects to improve implementation.	Analysis of net effects (treatment effects) of development outcomes to improve project design (concurrently or for future similar projects).
Data collection is ongoing, relying on readily accessible and regularly collected data.	Data collection is periodic, more intensive and requires information on both beneficiaries and non- beneficiaries over time.

6 Treatment effect methodological standard depends critically on the capacity for establishing counterfactuals. The highest standard for obtaining net effects is from a sampling strategy that uses randomization. Randomly allocating eligible beneficiaries into treatment and control groups draws them from the same population distribution, making the two groups perfectly comparable. This need not imply denying benefits to intended participants. Randomly allocating beneficiaries into first-round and second-round phases of implementation is sufficient. In the absence of randomization, two considerations come to the fore regarding sampling strategies. One is how to include comparable non-participants in a data set when the evaluator has no a priori way of knowing exactly who is comparable to the targeted beneficiaries. The other is what other variables need be to be accounted for, both to explain program participation, and to account for other influences on intended and unintended development outcomes.

7 "More detail should be given in defining what a "reflexive" evaluation is…it is not clear why reflexive and impact evaluations would have differential contributions to the effectiveness of an operation."

2.6 Note that the treatment effect method is not new to the Bank. The standard adopted by OVE is based on quality rigorous evaluations that have been carried out jointly by the Bank's professional staff and country authorities of IDB projects (see Box 2.2). The problem is not only that their paucity of numbers suggests that they are an exception to normal practice but also that they are exclusively evaluations of social projects. Given the lack of links between these evaluations and the Bank's routine information systems, there also appears to be insufficient dissemination of their findings, which are unavailable on Bank web pages or publicly available IDB documents.

Box 2.2: Examples of Quality Self Evaluations of IDB Projects

- Opportunities/ conditional cash transfer program/ Mexico/ quasi- experimental design evaluation. (See http://www.iadb.org/ove/DefaultNoCache.aspx?Action=WUCPublications@ImpactEvaluations)
- PRAF/ conditional cash transfer program/ Honduras/ experimental design. (See http://www.iadb.org/ove/DefaultNoCache.aspx? Action=WUCPublications@ImpactEvaluations)
- Social Network/ conditional cash transfer program/ Nicaragua/ experimental design. (See http://www.iadb.org/ove/DefaultNoCache.aspx? Action=WUCPublications@ImpactEvaluations)
- Social Network/ Colombia/ Quasi-experimental evaluation. The evaluations includes the following sub-programs: (i) Families in Action (conditional cash transfer); (ii) /Workfare; and (iii) Youth in Action (labour training)

2.7 Given the evaluative questions posed, the treatment effect method is able to estimate the impact of a treatment of an existing program and is the most appropriate to use in the implementation of this set of EPP. However, it has disadvantages (see Box 2.3). First, the range of questions that can be addressed is limited to an existing program; it cannot predict the effects of a new program or the same program in another context. Second, the ranges of programs that can be evaluated are in general limited to those of only partial coverage, so that treatment and control groups can be constructed; this usually excludes nation-wide programs. Finally, the findings of a single treatment effect evaluation may be of limited comparability across studies. To answer these different evaluative questions, a structural econometric approach would be more appropriate.

Box 2.3: Comparison of Alternative Approaches to Program Evaluation

	Structural Econometric Approach	Treatment Effect Approach
Range of questions addressed	Evaluates the treatment effect of existing program. Forecasts the program's effect in a new environment. Predicts the effects of a program never tried before.	Evaluates the treatment effect of existing program. Evaluates one program in one environment. Cannot predict effects of a new program.
Range of programs that can be evaluated	Programs with either partial or universal coverage depending on variation in data (prices and endowments	Programs with partial coverage (treatment and control groups)
Comparability Across studies	High comparability across evaluations (program invariant parameters)	Not generally comparable unless evaluations designed for a meta-evaluation of similar programs.

Source: modified from Table V in "Structural Equations, Treatment Effects And Econometric Policy Evaluations" by James J. Heckman and Edward Vytlacil, NBER Working Paper No. 11259, March 2005. Note this article proposes a synthesis of the two approaches, which is ignored in this modified table.

2.8 In addition, there are other evaluative questions, such as those regarding institutional sustainability or qualitative outcomes that require alternative methods of analysis. These questions do not lend themselves to either treatment or reflexive evaluations, given that (i) the dimensions that one would like to evaluate are not easily measured and (ii) the unit of analysis (institutions, etc.) are usually not numerous enough to apply standard statistical and econometric tools.

2.9 One issue that has come up in discussions with Management is the utilization of cost-benefit analysis versus of impact evaluations. There are three points to be made here. The first is regarding ex ante cost benefit analysis, or cost benefit analysis that are done prior to program implementation. Ex ante cost benefit analysis are one of a number of simulations that can be done prior to an intervention in order to obtain some idea of magnitude of benefits and costs[8]. They have been the mainstay ex ante simulation technique for the last thirty years or so, and to the Bank's credit, during the 1970s and 1980s it was a pioneer in applying this methodology as an ex ante criteria for investment eligibility[9]. However, although these simulations can help guide good policy, they cannot answer questions regarding the effectiveness of interventions; they can only provide ex ante "best guesses" of this effectiveness. This leads to the second point: cost-benefit analysis can be performed ex post also, as bonifide evaluation tools. However, in these cases they are not independent of impact evaluation; on the contrary, when done competently they will depend critically on the results of impact evaluations in order to parameterize the benefits side of the analysis. In other words, the parameters obtained from an impact evaluation should provide the building blocks of the benefits side of a cost-benefit analysis. Indeed, this is applicable to not only cost-benefit analysis but to any other type of evaluative analysis that attempts to compare the benefits of different interventions and hold them to a similar yardstick. The last point concerns the implicit assumptions of cost benefit analysis. Although cost benefit has the advantage that it identifies all costs and benefits of a policy or program and summarises them into one monetary measure of the aggregate change in welfare (hence provides a simple rule in decision makers of the desirability of a particular policy or program), it is neutral with respect to the distribution of costs and benefits. For the IDB this constitutes a problem given the institution's concern regarding poverty: one dollar for the poor is not the same as one dollar to a millionaire.

2.10 **Sampling Strategy.** OVE's selection of projects was purposive rather than random.[10] This reflects the choice between doing either evaluations of stand-alone projects and

[8] Recently there has been a proliferation of other simulation techniques aimed at obtaining projections of program benefits and the distribution of these benefits. See, for example, Bourguignon, Ferreira and Leite (2002).

[9] When comparing the practice of project evaluation that was done in the context of older Bank projects to the current evaluation practices done today, it has become evident that although the Bank was on the cutting edge of evaluation two decades ago, it has been unable to keep up with the evolution of technology and has since lost its comparative advantage.

[10] Other than the idea of clusters the selection criteria are in line with the suggests of EPP: The criteria to decide what projects should be subject to ex-post evaluation and the specific methodology to apply in each case include: i. OVE's priorities based on the areas targeted for evaluation. ii. The innovative aspects of the operation and its potential for drawing lesson learned. iii. The strategic relevance at the Bank, sector or national levels. iv The

simultaneously doing evaluations of clusters of similar projects (or sub-components of projects). A cluster evaluation allows for quantification of the importance of the project context vis-à-vis the project model in the realization of development outcomes, thus increasing the comparability across different evaluations. To the extent that the project context varies over time and across countries, looking at like projects over different countries and different years will hold the "model" constant and vary the "context". This implies evaluating similar projects simultaneously as "clusters" of projects. By "similar" projects it is meant projects that have a common model where model is defined by the commonality of the set of inputs, outputs and expected outcomes. Commonality thus allows their aggregation into a theme. In turn such an approach facilitates a meta-evaluation: an evaluation of the individual evaluations to obtain general lessons found. A meta-evaluation would have a greater value added in terms of knowledge generation then a single project evaluation to the extent findings are common across different implementation contexts.

2.11 The sampling strategy adopted the following criteria to determine which clusters of projects to select. First, relevance that is generally defined as choosing clusters of projects where there is a lack of evaluative knowledge and where there are many similar projects open and in the Bank's pipeline. However, in exceptional cases "stand-alone" evaluations of innovative projects could be selected if they are considered possible candidates for a future "cluster". Second, feasibility that is defined as clusters of projects in which there are many projects done or in execution as well as where there is a high probability that required information would be available. The latter criterion implied: (i) selecting projects whose expected development outcomes could probably be obtained from alternative sources (for example from population or agricultural census or household surveys etc.) in case no information was available in the Bank's Monitoring and Evaluation system; and (ii) selecting the most recent of projects that were approved or closed, i.e. limit projects to those approved since 1990. By doing so it was hoped to increase the likelihood of finding the actors involved in the projects and relevant documentation and data.

2.12 **Themes and Projects selected.** It should be noted that the 2004 cycle was *sui generis.* First, for the 2004 cycle, OVE avoided "programs" (i.e. operations with economy-wide effects) to ensure that projects evaluated had definable set of non-beneficiaries, thereby facilitating conventional treatment effect evaluation techniques.[11] Second, during 2003, OVE and Management met repeatedly to select the six pilot projects included in RE-284. These projects were incorporated into OVE's 2004 work program[12] that also included the proposal for ten additional evaluations, chosen within two clusters: Land Tenure Regularization and Neighbourhood Improvement. The dialogue with the Administration (which was represented by the development effectiveness coordinators of the three operational departments and staff member of DPP) produced more than just a consensual list of projects to be evaluated. Within the frame of a constructive dialogue the

---

particular interest and capacity of the Borrower, especially considering how the evaluation exercise may help to strengthen the country's evaluative capacity." page 3 of GN-2254-5

[11] This criterion implied avoiding policy based loans (traditional and emergency terms). The 2004 cycle also excluded technical assistance operations and private sector operations.

[12] These six had financial support of the Administration (Regions I and II plus DPP).

Administration brought to the table a collective experience that facilitated a clearer definition of an ex post evaluation road map in which many potential pitfalls were identified *a priori* as were optional routes in case such pitfalls occurred. Collaboration thus reduced information asymmetries and accelerated movement along the learning curve. Finally, as had been identified in the discussion with the Administration, juggling the number of projects' target simultaneously with the method target implied that some projects originally selected had to be dropped and new ones adopted during the exercise. For example, of the original six pilot studies, retrofitting of information was relatively successful for treatment effect evaluations for only three (BR-0182, EC-0025, and AR-0169).

2.13 Therefore, in early in 2004 three of the originally proposed evaluations were dropped: (i) Wawa Wasi, PE-0167, which was replaced with Land Titling and Registration, PE0037. This project was chosen for two reasons. First, it could be folded into the Land Titling Regularization cluster. Second, the Peruvian program distributed property titles to over 1.1 million rural households, making it one of the largest formalisation programs targeted to rural areas in the developing world. The program thus provides an opportunity to test the channels through which titles can effect household welfare; and (ii) ME –0186 was replaced, at mid-year, with PROCAMPO, ME-0185. The co-existence of PROCAMPO together with Mexico's PROGRESA program represented a unique opportunity for testing the need and role of conditionality in poverty alleviation cash transfer programs as there were households that were both receiving PROCAMPO transfers as well as receiving transfers from Mexico's conditional cash transfer program PROGRESA. It also allowed testing the oft-incorporated assumption that outcome benefits of cash transfers are maximised if the transfer is handed to the woman of the household (PROGRESA) rather than the man (PROCAMPO) independently of the gender of the head of the household. The last of the six pilot studies, the National Environment Project, ES0024, although begun in 2004, had its completion target date postponed to 2005 (as it required follow through primary data collection the cost of which was beyond the budget possibilities). It was replaced by PROAPS/REMEDIAR, AR—0120. REMEDIAR was chosen as a stand-alone evaluation because it is an example of an innovative project. It represents an example of how the IDB can react quickly to unanticipated circumstance of extreme economic crisis and to test whether the projects design was effective in mitigating the negative consequences of the crisis on the health status of the poor. While the criteria for dropping originally selected pilot projects related principally to lack of data and greater relevance and data availability of alternative projects, it required constant revisions to the agreements reached earlier with the Administration, a factor that should be taken into consideration in further rounds of ex post evaluation.

2.14 Table 1.1 summarises the list of projects that were originally and finally selected. In terms of the relevance criteria, at the moment of selection for Land Tenure Regularization projects, there were six closed, fourteen open, and five operations in the Bank's pipeline, while for Neighbourhood Improvement projects, there were nine closed (of which seven were selected), eight open, and ten in the pipeline. With respect to the stand-alone projects the Bank has approved during the nineties about: 67 Water and Sewerage projects; 18 Labour Training programs; 78 Social Investment Funds; and 6 Conditional

Cash Transfer programs. The table also highlights that most projects were a follow-up operation or had a follow-up operation: of the 16 selected 13 were follow-up operations or were followed by a follow-up operation, and where many follow-up operations were approved while the previous one was still open. Finally, under the interpretation described in chapter II, the projects selected implied attempting to carry out: (i) eight EPSA i.e. projects that closed at least two years but less than four before 2004; (ii) five IER i.e. projects that closed before 2000; and (iii) five "others" (including two projects that closed after 2000) and three that were still open. The reason for the latter category (outside the 2 or 4 year rule of EPP) was to test the hypothesis that treatment effect and reflexive methodologies could be applied to more recently closed or even open projects.

Table 1.1: Projects Selected for Ex Post Evaluation

Project	Title	Approval/ Closure Date	Program before	Details	Program after	Details
Land Tenure Regularization Projects						
PE-0037	Land Titling and Registration	1996/2001	no	-	yes	PE-0107 Register and Land Title Stage II (2001/2005)
PR-0083	Consolidation of Rural Settlements	1992/2000	no	-	no	-
EC-0048	Agricultural Sector Program	1994/2002	no	-	yes	EC-0191 Land Titling and Registration (2001/2007)
JA-0030	Land Titling Program I Stage	1987/1997	no	-	yes	JA-0050 Land Admin. & Management Program (1999/2005)
BL-0007	Belize Land Administration Project	1997/2001	no	-	yes	BL-0017 Land Administration II (2001/2006)
NI-0020	Agricultural Sector Program	1992/1997	no	-	no	-
Neighbourhood Improvement Projects						
*BR-0182**	*Rio de Janeiro Urban Upgrading Program*	*1996/2000*	*no*	-	*yes*	*BR-0250 Urban Improvement Rio de Janeiro II (2000/2006)*
EC-0138	Housing Sector Program	1997/2000	no	-	yes	EC-0207 Housing Sector Support Program II (2002/2006)
CH-0118	Sites and Services Improvement	1989/1994	yes	CH0134 Urban Infrastructure Program (1982/1988)	yes	CH-0032 Sites-Services and Neighbourhood Improvement Program (1994/1999)
UR-0070	Municipal Development Program II	1990/1998	yes	UR-0028 Municipal Development Program (1984/1992)	yes	UR-0111 Municipal Development Program III (1997/2004); UR-0131 Municipal Development and Management Program (2002/2006)
TT-0016	National Settlements Programs	1991/2001	no	-	no	-
Stand-Alone Projects						
*EC-0025**	*Water Supply and Sewerage Project for Quito*	*1994/2002*	*no*	-	*yes*	*EC-0200 Environmental Sanitation of Quito Metropolitan District (2002/2006)*
PN-0054	Social Investment Program	1994/2001	no	-	yes	PN-0111 Poverty Reduction and Community Development (1999/2005)
*AR-0120**	*PROAPS, Remediar*	*1999/2006*	*no*	-	*no*	-
*AR-0169**	*Youth Productivity Program*	*1997/2005*	*yes*	*AR-0062 Productive Recon version*	*no*	-
*ME-0186**	*Labor Market Modernization Project, Phase I*	*1996/2000*	*no*	-	*yes*	*ME-0118 Labor Market Modernization Project, Phase II (2000/2004)*
*ES0024**	*National Environment Project*	*1995/2006*	*no*	-	*no*	-
*PE0167**	*Wawa Wasi*	*1998/2005*	*no*	-	*no*	-
ME-0185	Food and Agricultural Sector Loan	1996/98	no		yes	ME0213 Procampo Support Program (2001/2005)

Italics identify the six pilot studies proposed in RE-284. ME-0186, ES0024 and PE0167 were those projects dropped during 2004 for lack of data, leaving the group of 16 ex post evaluations undertaken.

2.15 **Implementation.** The actual mechanics of implementation, once the projects were selected, was decided over the course of 2004 and early 2005, with frequent revisions of approach according to feasibility and other concerns. Implementation consisted of the following set of activities: First, a desk study was carried out in order to determine what was known about the interventions. A number of Bank sources were tapped: (i) loan documents and their accompanying documentation; (ii) ex post evaluations generated in

the Borrowers Ex Post sub-system; (iii) Project Completion Reports; (iv) Project Performance and Monitoring Reports; and (v) meta-evaluations made by different entities of the Bank. Second, OVE conducted interviews with project team members and COF specialist, and, where possible, country officials. These measures were aimed at identifying the relevant evaluative questions, sources of information on development outcomes, and list of contacts in the country, etc., to add to the information gathered in a desk review. This task was critical as project teams and local actors have information that is not picked up in the Bank's formal information system. Third, decisions were taken on whether to carry out the evaluation in-house or to (partially or wholly) outsource them, as well as which intra-institutional coordination mechanisms to use. Different institutional coordination mechanisms were adopted in different contexts; this variance allowed a first approximation at learning what "works best". Finally, quality control of individual evaluations was sought through technical seminars open to IDB staff and by subjecting individual reports to external technical reviewers.

2.16 OVE also has begun to put in place a number of mechanisms to build up a system that could assist in the task of doing a large number of project evaluations in the future. Three networks are in the process of being consolidated and will be used in the 2005 cycle:

(i) IntraEvalNet of IDB staff. This network's objectives are to bring together Bank staff interested in the evaluative agenda, and to interchange experiences, lessons learnt, terms of reference, techniques, etc., and thereby improve evaluative capacity.

(ii) EvalNet of program evaluation professionals in LAC. By building up a base of increasingly experienced evaluators in LAC, the objectives of this network are to construct a consultant roster to facilitate contracting out evaluation work by OVE and by the Bank. It will also enhance in country non–government evaluative capacity by working with local evaluation specialists. As outsourcing of evaluative work is the norm for both the IDB as well as governments, there is a high potential return from this network.
As of April 2005, 267 evaluators had registered representing 26 countries. 230 of those registered are from borrowing member countries.

(iii) RedEval of public entities involved in evaluation. The objectives of RedEval are to establish relations with official country counterparts for OVE's evaluations; and to build up in-country institutional evaluative capacity. The "network" remains at an embryonic stage, and requires greater effort, time, and budget to become effective.[13]

13 The potential benefits can be illustrated by the first conference of public evaluative agencies that was held in Brasilia in January 2004, organised by IPEA of Brazil, where representatives of over eight countries participated and interchanged experiences of different evaluative systems. A key conclusion of that meeting was the need for a network of public entities involved in evaluation.

III. PROJECT EVALUATION FINDINGS

3.1 The purpose of this chapter is to present the office's findings from the sample of projects that were selected for the 2004 cycle. We briefly present the findings by theme and also highlight some individual projects' evaluations[14] and end the chapter on the cost of the 2004 exercise. Full reports are available on IntraEvalNet: (http://ove/oveIntranet/Default.aspx? Action=WUCHtmlContent@EXE).

A. Thematic Review: Land Regularization Projects

3.2 An understanding of the Land Regularization projects is enhanced when it is viewed as part of a general development policy shift. During the nineties, policy in Latin America shifted towards a pro-market regime in which property rights were seen as a critical prerequisite for successful economic development and poverty reduction. In particular, legal land titles were considered critical for poor small farmers to benefit from markets. It was assumed that legalized land rights would reduce the risk of expropriation, which was considered to act like a tax on the returns of investment hence leading to underinvestment. Moreover, improvements in land rights were expected to reduce transaction costs and therefore also increase the market for land. Similarly, assuming competitive credit markets, land titles could be used as collateral reducing the risk premium on lending hence would reduce the restriction to access and interest faced by poor rural borrowers. Land Tenure Regularization projects had a common underlying model. In terms of outputs of the seven projects reviewed, five had institutional strengthening of administration of land records management and five identified land titles issued or registered. In terms of expected outcomes, seven had better land security, increase in productive investment in land, and five had more efficient, transparent land markets. Thus this theme's evaluations attempted to determine the treatment effect of land titles on investment in the household or plot, trade in land, and credit access.

3.3 None of the projects reviewed in this category had a complete set of development outcome indicators defined, nor was adequate information collected and maintained by the Bank in its monitoring and evaluation system. Neither was outcome data successfully obtained from the executing agency or project implementation unit. Six of the projects mention ex post evaluations in loan proposals, but none have been completed to date. OVE was successful in retrofitting a subset of outcomes expected for three projects: an attrition rate of 50%. This reality limits the information available regarding development effectiveness, hence provides insufficient information for a quality meta-evaluation.

3.4 The land regularization evaluations show an ambiguous treatment effect on productive investment. These ranged from a null treatment effect in Paraguay to a substantial and positive treatment effect in the case of Nicaragua. The treatment effect on access to credit was likewise ambiguous, with no demonstrable increase in access in either Nicaragua or Peru. The treatment effects on rental markets were found to be positive only in the case of Nicaragua, as was the treatment effect on agricultural productivity.

[14] A more detailed exposition can be found in Annex 1.

The only treatment effect that was large and unambiguous was the case of property values. Beneficiaries of Land Regularization projects saw property values for their land increase in all three projects. However, for the other purported development effects (greater productivity, increased investment, and greater access to credit), no unambiguous treatment effects were found.

3.5 The findings suggest that such programs cannot expect that just having a title will imply that formal credit institutions will automatically accept them as collateral, particularly in contexts of credit rationing and where informational asymmetries dominate the micro-credit environment. Specifically, smaller farms (less than twenty hectares) remain rationed out of the credit market and do not benefit from credit supply effects of a title. This differential effect between small and large properties, suggests that in environments where markets (like formal credit supply) entail distortions, smaller and poorer farmers are at a disadvantage, and may, given the direction of land sales towards larger sized producers, imply increased inequity of property. For small and poor producers to benefit from a pro-market regime requires that together with titling, transaction costs and market distortions that limit access to credit must be simultaneously reduced.

B. Thematic Review: Neighbourhood Improvement Projects

3.6 Neighbourhood projects based on the early experiences in Chile and later in Brazil have been replicated (or are in the pipeline) in most borrowing countries. However, an assessment of the components and objectives of these projects reveals surprisingly little commonality. The list of expected outcomes varies considerably from one project to the next. The objective that is most often cited is that of providing public services (water, sewerage, etc.), which is in fact an output of most projects rather than an outcome.[15] Other common objectives were to (i) develop housing markets, (ii) integrate neighbourhoods with the city and (iii) improve social indicators. In terms of outputs, there are two components found in all projects (i) public works and (ii) institutional strengthening, although there is

Box 3.1: Results from interviews of NIP team members

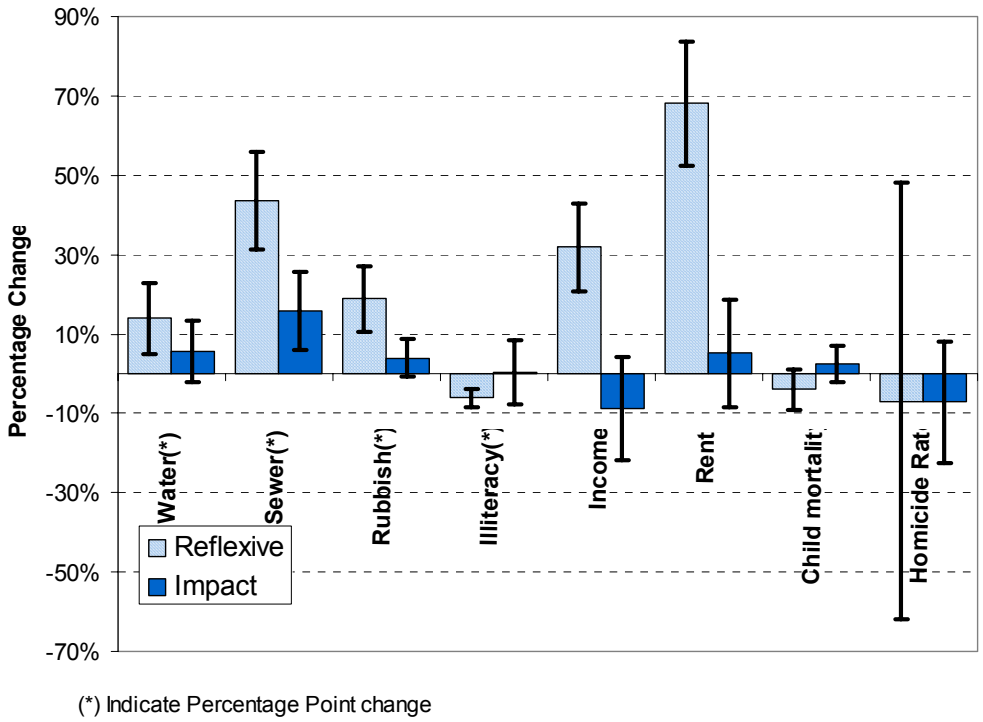
OVE's interview of team members revealed the heterogeneity of opinion regarding what, precisely, neighborhood improvement projects are designed to accomplish. In some cases interviewees mentioned that the main goal was to integrate neighborhoods with the city, and that this would be accomplished by reducing the differences in access and in legal rights of informal and formal city dwellers. In other cases respondents thought that the most important aspect was the reduction of transaction costs engendered by improving roads. In yet other cases mention was made of the potential of these interventions to affect social indicators, for example, morbidity and child development health indicators. In other cases attempts were made to link neighborhood improvement goals to the broad concept of capabilities developed by Sen. In this case the opinion given was that one should attempt to help informal neighborhoods attain the goals that they set internally; and so that for this reason different projects would have different goals. On the other hand, other interviewees expressed the opposite opinion. In one case the interviewee highlighted the dispersion of objectives across projects as a good example of "goal drift" within the institution (IDB).

[15] Many projects use unsatisfied basic needs (UBN) as an outcome measure, which simply measure the extent of access to different services (water, sanitation, garbage collection, electricity, etc.). However, there is a literature supporting the view that UBN are a reasonable approximation to well-being, and thus an outcome.

considerable variation in the nature of institutional components. Other components are wide ranging, and have included property rights titling, day care centers, health counselling, community driven investments and youth training, among others. The lack of a common rhetorical Output→Outcome→Goal framework suggests that, unlike land titling projects, there were no common underlying models in neighbourhood improvement projects. Interviews with NIP team members (see Box 6 which paraphrases the interviewees) corroborated this view. However, their common design feature is multi-dimensional integrated interventions (multi-treatments), under the assumption that there is a synergy effect. This view is reflected in the Bank's social development strategy approved in 2004.

3.7 OVE was less successful at retrofitting neighbourhood improvement projects than land titling. The office was able to obtain pre- and post-program information on one project (BR) through the use of third-party sources (census, household surveys and health secretary data): an attrition rate of 80%. In at least one case (Chile), this approach could have been replicated, but budget constraints did not allow for it. In two cases (Ecuador and Uruguay), an attempt was made to identify the program's treatment effect by looking at expenditures by municipalities on neighbourhood and sanitation (Uruguay) and by looking at the extent of program coverage (Ecuador). In the case of Uruguay, the results suggest that additional sub-national spending on sanitation and neighbourhoods leads to decreases in Unmet Basic Needs—particularly sanitation-related UBN. This treatment effect was also found in the case of Brazil. This effect is furthermore well targeted among the poor, both in the case of Brazil and Uruguay. However, in the case of Ecuador, no such effect was found. Furthermore, in both Brazil and Uruguay, there was no demonstrable treatment effect of the project on an index of housing quality. The same is true in Ecuador. The contrast between reflexive and treatment effect estimates in Brazil is shown in Chart 3.1. The chart shows the extent to which development effectiveness can be over- and under-stated in a reflexive evaluation.

Chart 3.1: Rio Neighbourhood Project: Reflexive and treatment effect comparisons



3.8 The overall conclusion regarding the small sample of projects evaluated is that they were successful at providing the target populations with greater coverage of certain public services. To the extent that these services are considered merit goods, their provision constitutes a positive result. In addition, in two of the cases, this impact was more pronounced for the poorest segments included in the treated population. Beyond this, very little else can be said. The impact on the objectives related to human capital formation and income were not demonstrated. In the case of health interventions, perhaps the intervention type most directly linked to sanitation services; there has been no demonstrated link between the interventions and outcomes, even for the poorest segments of the beneficiary population. There was also no consistent evidence showing an increase in variables related to housing values in the three projects reviewed[16]. With respect to the impact of the titling component of the projects, an additional evaluation will have to be undertaken, given that these components had not disbursed. The paucity

[16] In the case of the Rio project, there are two studies that were done regarding housing values. In one case the results showed a substantial appreciation in housing values; in the other there was no such appreciation. OVE's review did not find an appreciation in housing values. In the case of Ecuador, housing characteristics were selected as a proxy of housing value; there was no improvement in the type of walls and floor, the type of ceiling, or the number of rooms. In Uruguay, the measure looked at was the number of people per room, and again there was no evidence that more construction was undertaken. Here one should point out that the impact parameter is based on municipal spending, which includes IDB and other funds.

of treatment effect evaluations conducted in this thematic area leaves open whether a "multiple treatment" has synergy effects or not.

C. Stand –Alone Evaluations

3.9 Beyond the thematic evaluation results summarized above, results from individual projects were also useful in highlighting critical dimensions of evaluation. The first is the link between the existing literature surrounding a model and the need to do further evaluation. In the case of Quito's Water evaluation, the evaluation indicates that the preponderance of evidence shows that increasing the extent and quality of water has an important treatment effect on health status. This finding from the literature would suggest that very little could be learned from additional evaluations looking at water and health status. However, even in this case, the results of the evaluation revealed surprising and potentially policy-critical findings. The results of the treatment effect evaluation indicate, as expected, a positive treatment effect of water supply and sewerage services on health outcomes proxied by child mortality (on average, 8.0 percent higher without it, this figure is half that obtained from reflexive evaluation of 16.5 percent). However, the heterogeneity of results shows those more educated mothers were better able to take advantage of the benefits of increased water coverage. This led to a regressive relationship between treatment effect and income, where more educated (and wealthier) households did better than less educated (and poorer) households. These results suggest that among the poorest households, the education of women is a decisive factor to obtain expected child health benefits from water and sewerage interventions. The potential implication is that future projects should include or be coordinated with, as a hypothesis to be tested, a health education component together with water expansion.

3.10 The Panama Social Investment Fund-SIF (PN-0054) highlights a second evaluative finding. It shows that in some cases the reflexive evaluation in fact understated the true program treatment effect. The development outcome of this project was the reduction in poverty**.** A reflexive evaluation (the gross effects) of the incidence of poverty suggested that not only was the project unsuccessful but that it actually contributed to worsening poverty; the opposite of its intent. However, a treatment effect evaluation (the net effect) that compared "similar municipalities" shows that municipalities benefiting from FIS funds had a significant decline in poverty relative to comparable municipalities that did not receive FIS financing; the project had clear positive development outcomes.

3.11 In some cases, the results of the treatment effect evaluation were critical in identifying characteristics of an optimal program re-design during execution. This is the case of the School Scholarship Program in Argentina, a typical conditional cash transfer program. The timely evaluation was able to provide input into the timing and amount of grants. The policy question was how many repeated interventions (grants) per year has the higher payoff or is the pay-off the same for the same number of annual beneficiaries without repetition, a key targeting question for a program with a limited budget. The estimated treatment effects showed the highest cost effective pay-off was repetition of scholarships for three years. Another key issue in conditional cash transfer poverty alleviation program design is whether such demand-side interventions success depends upon supply side investment in schools and program implementation. This question was also tackled in the evaluation. An attempt was made to estimate the contributions of

program management and school characteristics in explaining variation between schools in student outcomes. The findings were that institutional capacity and conditions for learning in the schools and better communication and execution of program procedures account for significant variation of outcomes of the program: supply matters.

3.12 Another controversial issue in conditional cash programs[17] similar to that described in the previous paragraph is the conditionality itself. In poor countries, with less developed infrastructure and institutions, the administrative cost of monitoring compliance can be high. If a cash transfer alone is sufficient to induce behavioural change through income effects, the cost saving from monitoring compliance could be translated into increased coverage and/or higher benefits levels for the poor. However, the hypothesis that conditionality adds value and is cost-effective has not yet been evaluated. Two programs in Mexico provide an opportunity to do so. The first is PROGRESA, a conditional cash transfer programs that is a model for similar programs in the rest of the region. The conditions for receipt of transfers by participant households include a minimum number of preventive care visits to health clinics, a minimum level of school attendance (~80%), and attendance of a monthly health education seminar. The second program is PROCAMPO, launched in 1994, which is the first and only agricultural support program in LAC that provides direct cash benefits to farmers to compensate for losses expected under NAFTA. It includes none of the conditions imposed in PROGRESA. Thus the existence amongst the same group of poor rural households of PROCAMPO households and PROGRESA households provides a unique opportunity to observe the treatment effect of these two programs and attempt to determine the need for and role of conditionality. The findings are that the overall treatment effect of both programs on total expenditure and expenditure on food is the same. However, investment expenditure on health and education is higher in the PROGRESA case. Conditionality does result in an effect over and above the income effect of the transfer regarding education and health. PROCAMPO on the other hand results in greater investment in productive activities.

D. Lessons Learned

3.13 A critical finding across all projects is the lack of correspondence between the reflexive estimates and the treatment effect estimates. In practically all cases, the estimates were different. The Panama Social Investment Fund evaluation illustrated the case where the reflexive evaluation led to the conclusion of failure while the treatment effect method showed success. The Rio Neighbourhood study shows that for just one project nearly all-possible differences between reflexive and treatment effect can be found (see Chart 2). Comparing a reflexive finding to a treatment effect finding, the following held: (i) an over-estimation of a positive effect for potable water, sewerage and rubbish; (ii) an underestimation of a positive effect for child mortality; (iii) a positive effect when the treatment effect was negative (for income and rent); (iv) a negative effect when the treatment effect was positive (for illiteracy reduction); (v) that they are the same (for homicide rate). The ramifications are obvious: outcome data on just beneficiaries will

[17] Which based on the success of Mexico's PROGRESA that started in 1997 and Brazil's Bolsa Escola started in 1994, now exist in Argentina, Brazil, Colombia, Ecuador, Honduras, Jamaica, Nicaragua and Uruguay. It has become the most influential poverty alleviation initiative in the Region.

probably give misleading information on the development effectiveness. In addition, it shows the falseness of the assumption that treatment effect evaluations always show negative results.

3.14 Technical seminars were held with Bank staff to discuss the preliminary results of each of the evaluations undertaken. These seminars were advertised to the Bank at large and did not always succeed in attracting the key personnel working in each thematic area. In addition, working papers were distributed to key staff for comment and critique. The issues raised in these exercises provide important insights into the design of future rounds of ex post evaluation. First, given diffuse responsibilities for programming, preparation and execution within the Bank, several participants felt that evaluation, as an accountability exercise for the project team was counterproductive. Further, the contrast between reflexive and treatment effect evaluations raised issues around the utility of performance monitoring as an evaluation tool; was performance monitoring useless given findings on reflexive versus treatment? The utility and policy relevance of analyzing both intended and unintended effects was raised; if the project did not include the unintended effects in the logical framework, did this imply that the evaluation findings were not valid or useful? Given OVE retrofitting of third party data, participants felt that, in some cases, inappropriate variables were measured for a given development objective, leading to unsurprising negative findings. Finally, participants questioned whether the analysis of closed projects that were not required to include the necessary outcomes and data at the time of approval was a cost-effective use of Bank resources. These issues raise concerns that OVE shares regarding the EPP, but also shows a need for a shift to demand-led and real-time evaluation. Given the lessons learnt from the 2004 dissemination efforts, the 2005 strategy will be modified to assure the participation of key personnel and take into account, where relevant, the questions raised in the technical seminars.

E. Status and Cost

3.15 EPP's background document referring to ex post evaluations asked for "…*an estimate of their costs range*" (Para.4.8). This remains problematic. This is so for a number of reasons: First, projects selected during 2004 were selected because data was found or because it was thought they could be obtained cheaply (i.e. drawn from existing surveys or population census and administrative data of the project etc. Second, there is an upfront search costs that are a waste if no adequate data for a quality evaluation is finally found. Third, budget data does not incorporate total expenditure of individual evaluations as other actors involved absorbed part of the costs.[18]

[18] Overall costs can be viewed not only as the cost of doing individual evaluations, but also as part of institutional capacity building, since institutional capacity building also requires money. Part of OVE's ex post evaluation strategy has been to shift, to the extent possible, expenditure on a given project evaluation to local actors, with developing institutional capacity in mind. However, enhancing in-country evaluative capacity through sharing individual one-off project evaluation experiences is of limited value, unless complemented by the implementation of the "Third Pillar" of the Bank's medium-term action plan to improve development effectiveness.

Table 3.1: Costs and Status of Project Evaluations

Project	Title	Status	Total Costs ($)	Administration funds
PE-0037	**Land Titling and Registration**	**Quality evaluation/Working Paper**	**91,544**	**62,841**
PR-0083	Consolidation of Rural Settlements	Technical Note, some information for outcomes	23,577	
EC-0048	Agricultural Sector Program	Technical Note, qualitative studies done	23,777	
JA-0030	Land Titling Program I Stage	abandoned	12,457	
BL-0007	Belize Land Administration Project	Technical Note	23,577	
NI-0020	Agricultural Sector Program	Technical Note, cost benefit calculations	24,577	
BR-0182*	***Rio de Janeiro Urban Upgrading Program***	***Quality evaluation/Working Paper***	***83,606***	***29453***
EC-0138	Housing Sector Program	Technical Note, could not match up beneficiaries	16,777	
CH-0118 & CH0032	Sites and Services Improvement	abandoned	12,777	
UR-0070 (UR-0111)	Municipal Development Program II	Technical Note	16,777	
TT-0016	National Settlements Programs	abandoned	8,777	
EC-0025*	***Water Supply and Sewerage Project for Quito *1***	***Quality evaluation/Working Paper***	***89,934***	***39503***
PN-0054	**Social Investment Program *2**	**Quality evaluation/Working Paper**	**27,777**	
AR-0120*	PROAPS, Remediar	Technical Note	50,337	24865
AR-0169*	***Schoolarship Program***	***Quality evaluation/Working Paper***	***37,545***	***12754***
ME-0186 & ME0118	*Labor Market Modernization Project, Phase I ***3***	*Technical note, meta evaluation*	*24,577*	
ES0024*	*National Environment Project*	*Postponed to 2005-2006*	*20,445*	
PE0167*	*Wawa Wasi*	*Not done, substituded by Schoolarship Program AR0169*	*9,777*	
ME-0185 & ME-0213	**Food and Agricultural Sector Loan & Procampo Support Program**	**Quality evaluation/Working Paper**	**50,000**	**18103**

The pilot studies are in italics. The quality evaluations are in bold. "Administrative funds" is the source other than OVE budget. Note of the Funds transferred from the Administration ($210,000) $187,515 was spent on six evaluations the balance returned to the common pool. The six included three originally selected as pilot studies and three that replaced the original ones selected.

3.16 Table 3.1 presents the individual project costs (staff, consultants and travel). Total expenditure was $648,615. The average cost per project was $34,000 with a high degree of variance (maximum was $91,541 and minimum was $8,777). However only six quality evaluations, i.e. those that satisfied the minimum method, were produced (with an average direct cost of $63,401).[19] About half the budget, $324,000 was spent on unsuccessful evaluation attempts. Incorporating total costs evenly into the six products implies an average cost of $113,694 per quality output. The costs per quality evaluation are on the lower side of ranges commonly advanced both in absolute dollar values as well as in percentages of the program's dollar value. The cost benchmarks are in terms of absolute dollars a range from $200,000 to 900,000 and as a percent of the program

[19] This is very similar to the estimated cost per project that was proposed in the pilot study: $70,000 per evaluation. Note that proposal was for $420,000 for six evaluations from the Bank's contingency fund. De facto a total of $210,000 was transferred from evenly the budgets of Region 1 and 3 and (as Region 2 claimed it could not find $70,000) the rest from DPP. The funds were made available to OVE in October 2003 two and half months prior to closure of use of funds on 15 December of 2003. The rest of the proposed cost $210,000 was absorbed by OVE in its own budget.

between five to seven percent. [20] The six-quality evaluations (see Table 3.2) ranged from $87,000 to $141,000 and as a percent of the dollar value of the programs (sum of follow-up operations) ranged from 0.01% to 0.21%.

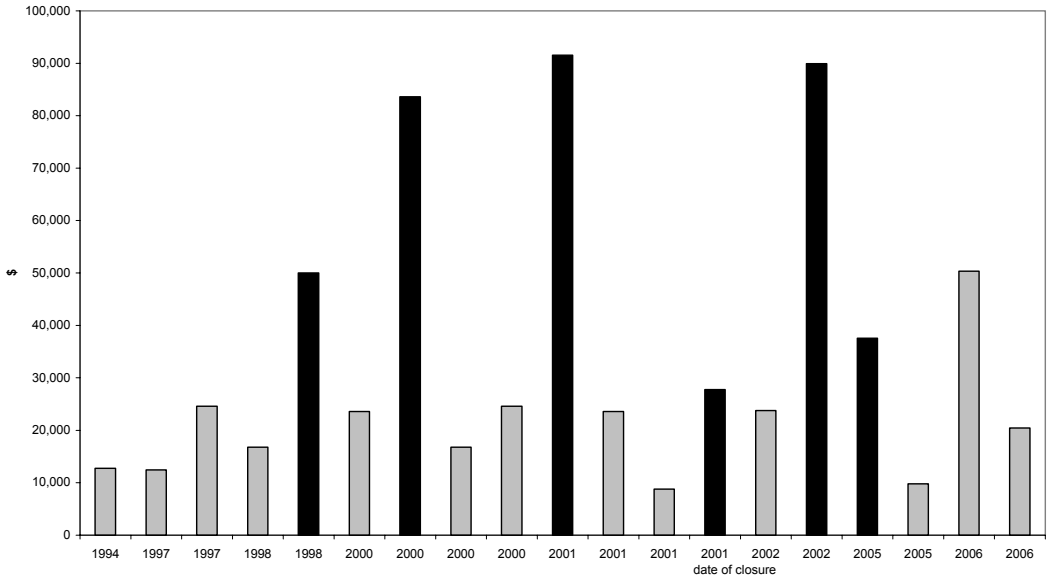
Table 3.2: Cost Summary of Quality Evaluations

	Program Value (US$ million)	Direct Costs (US$)	Total Costs (US$)	Total Costs as a Percent of Program Value
PE0037/PE0107	83	91,544	141,837	0.17%
BR-0107/BR-0250	600	83,606	133,899	0.02%
EC-0025/EC-0200	280	89,934	140,227	0.05%
AR-0169	637	37,545	87,838	0.01%
ME-0185/ME0169	1881	50,000	100,293	0.01%
PN-0054	38	27,777	78,070	0.21%

3.17 Finally, despite biasing selection in favour of high probable success rates the actual success rate in achieving quality evaluations was low (see Chart 3.2). Of the eight proposed EPSA, only 3 satisfied the minimum treatment effect standard. Of the five proposed IER, only one met the standard and of the five "others" (i.e. projects selected that did not satisfy the two and four year closed rule), two satisfied the standard. Further, no clear pattern of cost increasing the further back in time the project closed (particularly the assumption originally made that that an IER would cost at least double that of an EPSA) was found. However, the success rate of achieving a quality evaluation appears to be higher the more recent a project closed or if it is still open. The success of obtaining a quality evaluation for an open project suggests that treatment method can be applied to open projects.

[20] "There is considerable variation in the cost of impact evaluations. Information from OED suggests that impact evaluations are generally expensive, but can range anywhere between US$ 200,000 and US$ 900,000 depending on the use of methodology and extent of data collection (World Bank 2002, pg. 21). Another source suggests that the cost of an evaluation is between about 5 and 7 percent of the cost of the program being evaluated as a rule of thumb (W. K. Kellogg Foundation, 1998). From a sample of evaluations conducted in LAC, Judy Baker found that the average cost of the impact evaluations was US$ 433,000, but as a percentage of the loan or credit or total project cost, amounted to only about 0.56 percent. An informal survey of private firms that conduct impact evaluations found that a rigorous evaluation can be conducted in the range of US$ 300,000 – 350,000.7. " See John Blomquist, " Impact Evaluations of Social Programs: A policy Perspective", April 2003, http://www1.worldbank.org/sp/safetynets/Training_Events/ImpEval_Blomquist_Paper.pdf

Chart 3.2: Cost and Closure Dates of Projects Evaluated



Note "costs" are direct costs. Black columns represent the cases of quality evaluations.

IV. LESSONS LEARNED IN 2004

4.1 The purpose of this chapter is to draw lessons from the experience of the first year's implementation of EPP. As mentioned earlier, all 16 projects evaluated are considered pilots, as 2004 was the first year of implementation of the EPP. Specifically, RE-284 set out four operative questions on which more guidance was sought through the pilot which cover the topics of coordination, execution, cost, data, outcomes, effectiveness and lesson learning via dissemination.

4.2 Information available to the Bank at the time of designing the policy suggested that: (i) ex post coordination between Management, OVE and the country would resolve any problem encountered at the ex post evaluation stage; (ii) ex post evaluation could be an annual task that could start and finish within a calendar year; (iii) there would be sufficient evaluative information as an input to the ex post task; (iv) the costs for a treatment effect evaluation would be extremely large and would be complex to do, and therefore the major effort of the ex post task should be to concentrate on reflexive evaluations that were expected to be relatively cheap and simple to do; (iv) data on development outcomes (before-during and the end of project) would generally be available, and retrofitting would not be normally required, and that the ex post evaluation could draw on a number of evaluations done by the Bank during execution. The first year's experience suggests that information then available may resulted in excessively optimistic expectations.

4.3 **Coordination.** The information available in 2003 that was used as an input into the EPP suggested that ex post coordination would be both necessary and sufficient to resolve

problems encountered in the ex post task. However, coordination between Management, OVE and borrowing countries after projects close is too late to obtain satisfactory information and data for the ex post task. If data is not collected during preparation and execution, then retrofitting is required with all its problems (see Para 4.8). Despite these constraints, the quality of the evaluation is enhanced to the degree that intra-IDB coordination exists, i.e., staff that had been/are involved in the project being evaluated are actively involved in the ex post evaluation. Given the low level of reporting and record keeping (see below), coordination between the Administration and the Borrower in the shared responsibility for data collection and analysis is not working to make available quality information on development outcomes of projects.

4.4 **Execution.** The information available in 2003 that was used as an input in EPP suggested that an annual ex post evaluation exercise would be feasible within a calendar year. However, the initial timeline set for the pilot, and indeed for the 10 additional evaluations financed through the OVE work plan, was unrealistic. Even those pilot evaluations that were thought to have sufficient data as a result of a desk review, once analyzed by evaluators, these data were insufficient to meet the methodological standard set and thus required retrofitting. Further, future efforts that include more intensive original data collection will be more time-consuming. The 2004 experience, with its high attrition rate that will likely be typical of future efforts to evaluate the stock of projects, demonstrates the extent of preparatory work required for the EPP as currently conceived.

4.5 **Cost.** The information available in 2003 that was used as an input in the EPP suggested that treatment effect evaluations would be very costly, about $250,00 while reflexive evaluations would be much cheaper, about $30,000. However, with the many caveats cited in chapter III, ex post, or indeed concurrent, evaluations appear costly in budgetary terms (approximately 20% of OVE total budget in 2004), although much lower than expected in terms of the magnitude of the original lending (less than half a percent) or in dollar terms, average $63,401, with no significant difference between reflexive and treatment effect standards. The 2004 experience illustrates the wide variation in costs. Evaluations that involve original data collection are clearly costlier than those that rely on existing data sources or add on to a routine statistical information gathering effort. Regarding complexity, if data is available there is no significant difference between reflexive and treatment effect methods. Also it must be noted that the Bank and Country normally jointly decide the evaluative strategy, agree upon the terms of reference of the evaluation and outsource the evaluation. The 2004 experience plus the quality of the individuals and entities in EVALNET suggest that most countries can draw upon professionals in universities, and/or profit and non-profit entities where they are located.[21]

[21] Amongst the Bank's borrowers three distinct institutional models of evaluation appear to be emerging, where: (i) evaluation is integrated into the country's budgetary system; (ii) a specialised evaluation office is in charge of coordinating evaluations; and (iii) Congress through a specially created commission has taken on the task. The velocity of this change amongst it borrowers poses the danger that the Bank's own practice falls behind those of some of its borrowers. An oft-reported theme in the interviews underlying this report is that while the Bank had high technical standards and own capacity during the seventies and eighties, although perhaps too high relative to the capacity of the Borrowers, today the situation is rapidly being reversed: it is probably lagging behind.

4.6 **Data: Evaluative Information.** The information available in 2003 that was used as an input to the EPP assumed that an ex post evaluation could build on existing evaluation design and information. This information was simply unavailable for the projects reviewed. Of the sixteen projects evaluated in 2004, although six had an evaluation strategy identified in the approval stage, most had abandoned the strategy during execution prior to project closure and, with one exception which produced data that could be used to calculate a treatment effect, none had produced quality evaluative information. In the cases of abandonment, the decision lay with the borrower. Although all projects reviewed completed a PCR, no PCR provided adequate information regarding the evolution of development outcomes expected from the project or an update with respect to the evaluation identified at the time of approval. Finally, regarding sector work produced by the Bank, no quality meta-evaluation was found regarding the thematic clusters selected or in the areas of the stand-alone projects. Quality meta-evaluation is defined as narrative based on the findings derived from treatment effect evaluations of a number of similar IDB or external projects.[22]

4.7 **Data: Data on outcomes.** Information available in 2003 that was used as an input for EPP suggested that data on outcomes required for a reflexive ex post evaluation was generally generated from design to execution to closure of all projects. However, of the sixteen projects evaluated, none reported information on outcomes of all their declared expected development outcomes. The general lesson is that the Bank does not have a comprehensive reliable data collection system for development outcomes of its projects. This reality is at odds with the Bank's declared objective of managing for results for evidence-based decisions to improve development effectiveness. In addition, OVE's experience in 2004 also indicates that data stored outside the IDB is relatively difficult to obtain and, in some cases, is inaccessible or lost. The immediate consequence of this information shortfall is that OVE has to attempt to retrofit information back to a time before the project started, and to when it closed (tasks of the Administration), in addition to the task set for OVE by EPP, which is to collect information at some point in time after project closure (i.e. in terms of Chart 1 points C back to B back to A).

4.8 **Data: Retrofitting.** Information available in 2003 implied that the need for retrofitting would be the exception rather than the rule. The opposite was found to hold. To fill the information gap was sometimes found to be possible. Of the sixteen projects, seven projects were successfully retrofitted for *some* development outcome indicator for at least a reflexive evaluation and six for a treatment effect evaluation. However, retrofitting is not only a cost not anticipated in the EPP, it also reduces the degree of confidence in the applicability or relevance of the ex post evaluation findings, as data constrains the range of topics covered and methods employed. Specifically the problems are: First, projects do not provide unambiguous metric statements of the projects expected development effects. This implies that in an ex post evaluation the evaluator has to retrofit a metric, creating a potential source of conflict as to what the project designers "meant" as the development

[22] A meta-analysis of social investment funds' evaluations in the region, several of which were financed by the Bank, are reported in a publication of the World Bank.

objective.[23] An ex-ante definition by the project team of acceptable indicators of expected results can help establish the parameters for evaluating the development effectiveness of a project. However, it is important not only to set goals in terms of outputs and outcomes, but also to identify, at the project design stage, when a given goal is unattainable or the program needs to be redesigned during execution.[24] Second, retrofitting using existing secondary sources severely limits the feasibility of analyzing the continuity of benefits: for example, census data is produced every ten years. Third, the use of secondary retrofitted data means that the questions asked and answered in the evaluations were more dependent on the information found than on the relevance and usefulness of the hypotheses being tested: the tail wagging the dog. Fourth, retrofitted data may imply that the evaluative questions do not necessarily match with the specific development outcomes declared by the projects. A project can be evaluated using intended and unintended effects, but should at least consider as a minimum the intended ones. Retrofitting data reduces the probability of successfully doing so. Fifth, retrofitted data may not correspond exactly with the project's beneficiaries [25] or retrofitted data may not be able to identify the beneficiaries. [26] Sixth, it limits the set of control variables available and thus may reduce the reliability of treatment effect findings.

[23] In the case of EC-0138, the goal of the program was "to improve housing conditions for low-income groups." However, no indicators were defined to measure these conditions. In this case, based on the information available in national surveys, OVE analyzed housing conditions based on the following indicators: construction materials of floors, walls and roof, water connection inside the dwelling, existence of a shower and of a toilet, and the existence of an exclusive bedroom. The HQ specialist agreed on these indicators, but suggested that the program achieved other equally important objectives related to the purposes of the program, "make public housing expenditure more efficient and equitable, and provide incentives for commercial financial institutions and the private building industry to become involved in the housing sector", for which no indicators appear in the loan document and for which the specialist did not have specific suggestions. Another example is the Rio de Janeiro Urban Upgrading Program (BR-0182). This project had as an intended outcome to "*reduce the incidence of vector-borne diseases*" (PR-2077, 2.2), a precise and measurable outcome indicator. However, because information was not collected on disease incidence, OVE used data on mortality due to poor sanitation as a best-available proxy indicator. An objective of the Quito Water Supply and Sewerage Project (EC-0025) was to "*improve the hygienic and health conditions of the population of the city of Quito*" (PR-1984, 2.1). In this case, the indicator is imprecise, and can be mapped to any number of specific metrics. Since no measurable indicators were defined and no data was collected, OVE retrofitted using census data, and used child mortality as the best-available proxy indicator.

[24] In the active labor market programs in Mexico, a declared goal was to increase placement rates for the unemployed from 30% to 45% in the case of those receiving orientation on the job search process, and from 60% to 80% for those receiving training. Repeatedly, these goals are not met. The implication, however, is not on precise recommendations on how to adjust the programs, but rather that more money is needed; the underachievement was taken as evidence that the program should be expanded.

[25] For example, for the project PR-0083 (*Proyecto de Consolidación de Colonias Rurales*), OVE was able to obtain panel data on 300 households for the years 1994, 1999, and 2002 from academics in Universities of Wisconsin, Washington and Berkeley. Although it contains detailed information on the development objectives of the project, it does not cover the project's targeted part of the country, casting doubt on the relevance of the findings to the IDB project.

[26] For example, in the case of UR-0111, all rural provinces received financing. The information available does not identify beneficiaries within the province. One way to mitigate this limitation would be to obtain data at a more disaggregated geographical level. Micro information could have been reconstructed, with the assistance from the country's statistical institute, but budgetary restrictions in 2004 prevented this option. Budgetary restrictions also underlie the failure in the case of the projects CH-0118 and CH-0032. In this case, as in the case above, the data are only readily available at a very aggregate level. It is likely that the government would be able to produce richer and more disaggregated data, but the financing required to obtain and transform the data into a usable form was not available.

4.9 The picture formed by the findings regarding the data reported by the Bank on the development effectiveness of its projects is based on the sample used in the 2004 ex post exercise. The findings may be dependent on the sample and may not be generalisable. However, an exercise carried out on all active projects as of July 2004 revealed that data gap found in the 2004 sample generally holds for the Bank's entire portfolio: Of the 522 active projects, only 99 (18%) reconfirmed that data was being collected for at least one development objective of the project (see Box 4.1).

Box 4.1: Self Reporting by the Bank on Development Indicator Information of its Active Projects

Development Outcome Information. To determine compliance with the policy's promise of generating information to satisfy at least the reflexive standard, a census of active projects as of July 2004 was carried out. Self-reporting by the Bank shows that the minimum information standard it has set for itself is being met by a significant number of projects, in that their PPMRs affirm that baseline information exists, data collection system is in place and further data collection is occurring: 33% (i.e. 177 out of 522), although not the 100% promised by the EPP. A survey of these projects 178 projects reveals a different picture.

Of the 177 active projects, the survey (response was for 129) showed that only 99 (18% of active projects) projects re-confirm the existence of the information for *at least one development outcome indicator* but zero percent for all their declared development outcomes.

To gauge the accessibility of the data, the physical location of the information was requested. Of the 99 projects, the data's location is distributed as follows: 7 in COF (1%of active projects); 85 in Executing Agency (16% of active projects); and 17 in others, including 3 unknown. This will cause problems. Unless data is requested and obtained during project execution, the probability of obtaining the information after project closure is small.

Another dimension of availability was whether data was in electronic form. Projects with data in electronic format numbered 39.

The third dimension was which evaluation method standards could be used from the information gathered: 59 (11% of active projects) reported a reflexive standard (5 in COF) and 34 (7% of active projects) a treatment effect standard (3 in COF).

Regarding financing of the data collection system, 45% were financed from a TC or component of the loan, 40% from counterpart funds and the rest "other".

It is often asserted that the more recently approved the project, the more probable that an evaluative strategy exists, and that there is successful data collection during execution. The 99 projects by approval year do show an increase during the nineties but reach a peak in 2000 and thereafter show a drop. The figures represent upper bounds, as they do not consider the quality of the information gathered, and assume that information gathering is not abandoned during execution of the projects.

4.10 Further, although country statistical institutes have the capacity to design and collect information for individual projects' treatment effect evaluations for a reasonable price[27], according to a recent survey by OVE[28], no statistical agency in the region had been involved in generating data for a specific IDB project evaluation.

[27] Statistics institutes indicated that incorporating questions in existing surveys, modules in existing surveys or special ad hoc surveys would range from US$15,000 to US$30,000.

[28] Luis Marcano and Luis Miguel Lavista. Data availability of IDB projects: the findings of two surveys, OVE, 2005.

4.11 **Effectiveness: measurement of effectiveness: reflexive versus impact/treatment effect evaluations.** A reflexive evaluation may give misleading findings regarding either the size and direction or both of the development effects. The evaluations carried out in the 2004 cycle reveal that in very few cases were the findings regarding development effect between reflexive and treatment effect results equivalent. As the direction and magnitude of bias is usually unknown a priori, the minimum method has to be one that produces a treatment effect calculation whenever possible. While full information is not available on costs, the 2004 experience indicates that this method can be affordable.

4.12 However, as noted in Chapter II, treatment effect evaluation responds to one set of evaluative questions that in the 2004 exercise has focused on the effect of a program on its development goals. However, other types of analyses, such as analyses of benefit incidence, the relation between benefits and costs, and other intermediate outcomes are also extremely useful to program designers and planners.

4.13 **Lesson learning, products and dissemination.** The 2004 ex post cycle produced 14 written products and six seminars:

(i) Six Working Papers prepared for evaluations that met the minimum of treatment effect standard and passed the external review system. In terms of the timing definitions of IER and EPSA, these evaluations represent two IER and three EPSA. The sixth paper is a concurrent evaluation of an open project (AR-0169).

(ii) Eight technical notes prepared for evaluations that did not meet the minimum treatment effect evaluation standard. This product was not included in the EPP; however, it represents a best attempt at documenting whatever development effectiveness data was available with respect to the project under evaluation. Five notes refer to projects closed more than 4 years prior to the evaluation date, two to projects closed 2-4 years prior and one to an open project.

(iii) Six technical internal IDB seminars to discuss the Working Papers.

4.14 However, a task that remains pending is the dissemination of findings among borrowing countries. Efforts were made to assure that relevant ministries and executing agencies received copies of all documents produced. In addition, once the review process is completed within the Bank, all documents will be posted on the Bank's website. RE-284 mentions that a conference will be held to report on the 2004 cycle. This conference will be organized together with Management or alternative dissemination activities will be established.

V. CONCLUSIONS AND RECOMMENDATIONS

5.1 The 2004 cycle's conclusions can be summarised into six main points:

- The key to successful ex post evaluation is quality evaluative design at approval and data collection during project implementation. That project's development outcomes are defined during preparation and data generated during execution was found to be the exception rather than the rule, with the result of forcing a retrofitted evaluation. Neither a reflexive nor a treatment effect evaluation could be carried out based on the Bank's existing store of data.[29] A retrofitted evaluation is possible but problematic. While better than nothing, it increases evaluation costs and evaluates based on available third party data rather than data that most accurately reflect the development objectives of a project.

- To answer the evaluative questions posed in the EPP, reflexive evaluations will probably generate misleading findings. Treatment effect evaluation that establishes causal links between actions and outcomes may be considered the desired standard for partial coverage public projects, complemented by analyses of, costs, benefit incidence and institutional effects, among others. The costs of such evaluations depend on the aspects mentioned above, but are marginal with respect to the magnitude of the lending.

- Stand-alone project evaluations findings are difficult to assess; it is difficult to separate effects that are idiosyncratic to that project or its context versus the systematic result of project design. This problem can be addressed by evaluating a cluster of similar projects.

- The success of the results agenda and the role of ex post evaluation within it depends on borrowing country development of their own results management and evaluation agenda, and not only the Bank's internal capacity. The existing system of "shared responsibility" is not working well neither for concurrent monitoring nor for ex post evaluability of a project.[30]

[29] If the recent initiatives to implement the EPP vision regarding evaluability are effectively implemented, given an average disbursement period of ten years; the "real time" system promised by the EPP will be fully functioning in 2014.

[30] Few IDB projects start with quality evaluations as part of the original design, fewer are carried out, and almost none are available to the public. While some observers indicate that the borrower is responsible for the lack of an evaluation during design or for the abandonment during project execution, the evaluation results in the 2004 exercise were primarily based on information gathered through the borrower's permanent information systems. Very few results hinged upon information contained in the Bank's information systems. Indeed, whereas borrower systems contributed most of the quantitative data on development results, the main contribution of the Bank's systems—in the 2004 cycle—was either quantitative information on outputs or anecdote and opinion regarding development results.

- Better and early collaboration between Management, OVE and borrowers in the planning and execution of evaluations leads to better quality, more relevant evaluations. Further, the value added of concurrent and collaborative evaluation suggests that the EPP timing requirements with regard to IER and EPSA may be unnecessary.

- The six treatment effect evaluations undertaken during 2004 do show that the Bank's interventions have a significant development effect for at least one declared development objective. These findings suggest that the Bank may currently be understating its contribution to development.

5.2 During the 2005 cycle issues that were not tackled front and center will be addressed, such as (i) the applicability of cost-benefit analysis, (ii) greater emphasis on institutional and sustainability analysis in the EPSAs for 2005, (iii) analyzing the relevance of misleading impact estimates for the Bank and for the country. The following tentative recommendations make suggestions for the adjustment of the Bank's evaluation practice (concurrent to the project or ex post). They are tentative because OVE believes that more information, to be obtained from at least an additional year's experience, is required before a definitive set of recommendations can be addressed.

5.3 To answer the evaluative questions currently posed in the EPP, for both IER and EPSA, the minimum standard for evaluation of projects with partial coverage is probably a treatment effect evaluation. The veracity of this assertion will be checked for in the 2005 cycle. Timing requirements for evaluation should probably be eliminated and the evaluation products better differentiated in terms of the evaluative questions to be answered. Evaluation standards for economy-wide projects must be developed in future evaluation cycles, but will likely rely on structural econometric techniques (see 2.7) and other methods including cost-benefit analysis.

5.4 The 2005 cycle should focus its evaluation efforts on a cluster of similar projects, with greatest relevance to pipeline and portfolio. Projects have been selected that fit into three themes: (i) Job Training Programs, (ii) Rural Roads, and (iii) Science and Technology (see Annex 3 for a list and brief description of the projects). To date, the proposal includes 48 projects: 11 labour training (with Argentina and Mexico having back to back operations); 16 rural roads (Ecuador, Salvador, Guatemala, Honduras, Nicaragua and Peru having back to back operations); and 19 science and technology (with Argentina, Brazil, Chile, Colombia, Mexico, Uruguay, and Venezuela having back to back operations). The list of projects covers 16 countries. To obtain the number of evaluations required for a meta-evaluation, projects are over-sampled at the start of the evaluation cycle.[31]

[31] The 2005 cycle will over-sample projects in each cluster; 48 projects are included in the first round list. This strategy follows from findings in the 2004 cycle: (i) low probability in encountering adequate data in the Bank's Monitoring and Evaluation system; (ii) low successful retrofitting rate from secondary sources; and (iii) credible development result attributable to the project requires at a minimum an treatment effect evaluation. Taken together they imply the need for an over-sampling strategy to meet both the targeted number (about fifteen or so per year individual projects evaluated) and where the minimum standard is one in which treatment effect and reflexive estimations can be made. An attrition rate of about sixty-five percent is expected.

5.5 Greater and more formalised institutional coordination should be put in place to (i) carry out ex post evaluations of the existing stock of IDB projects and (ii) improve the probability of quality ex post evaluations in the future. For the 2005 cycle, Thematic Advisory Groups have been formed and are drawn from Management-nominated Bank professionals that participated in the projects under review or are the relevant specialist in the theme. Closer collaboration with Bank staff is hoped to improve the relevance of the evaluative questions formulated and the ownership of results. Borrowers, through executing agencies, evaluation agencies, finance and planning ministries, will also participate in 2005 evaluations. Incorporating country authorities also has the objective of increasing the ownership of the evaluation. The development of working relationships with country authorities will overtime be formalized as the Bank's PRODEV initiative, which has as one of its tasks of enhancing "… *planning and evaluation of investment projects*…" (Para 9)[32] comes on stream.

5.6 To improve the probability of quality ex post evaluation in the future, OVE will begin to evaluate the quality of evaluative design on approval and data collection during project implementation in the project. In 2005, OVE will design and implement an information system, jointly with the Administration, geared towards fulfilling the EPP suggestion that " *In the case of …impact evaluations, OVE will attempt to identify operations that will be subject to this type of evaluations as early as possible during the project cycle (i.e. during programming and design of the operation).*"(Para. 3.2).

5.7 The ultimate success of managing for results hinges on the adoption of evidence-based decision-making model amongst its borrowers. This will not occur overnight. Meanwhile the Bank's own system reveals a gap that is inversely proportional to its strategic ambition, which requires immediate steps to correct. Under this light the following three recommendations are made.

5.8 First, adopt ex ante evaluation as a routine activity of project preparation and treatment effect evaluation as the preferred method. Metrics must be clearly identified for the project's anticipated development outcomes and filling in the project's logical framework with numbers is necessary but insufficient. The Bank should incorporate ex ante evaluation during the preparation phase and ensure that an evaluation strategy to generate appropriate data is designed during approval of the project and that data is collected during execution. The Bank should explore as the norm the inclusion of the country's statistical system as early on as possible and to facilitate information gathering into the country's statistical system consider the benefits of giving a blanket procurement exemption to allow direct contracting by executing agencies of statistical institutes in country.

5.9 Second, rethink the institutional basis for generating adequate data for treatment effect evaluations of all its projects during execution. Recent efforts have concentrated on improving the design of monitoring reports and tacking onto the existing system the new information requirements. Results so far have been disappointing. This is to be expected;

[32] See GN-2346-2

the existing system was designed for and in practice remains a system to generate information (financial and on outputs) required for performance and process evaluations. The development effectiveness agenda creates data requirements that are both significantly different and substantially broader. Data collection is less frequent, more extensive (both beneficiaries and non beneficiaries and control variables) and requires both a completely different skill mix and counterpart institutional structure to that of the existing performance monitoring system.

5.10 Third, the Bank needs to formally define and institutionalise mechanisms that ensure the mapping from lessons found from the ex post evaluations to lesson learnt i.e. the institutional basis for evidence based decision-making.

5.11 OVE suggests that during 2005 the Management and the Office work together to produce an action plan that sets out a road map to bring together its strategic ambition (with detailed description of method standards, required skill mix, the institutional basis and the budgetary requirements) and actual institutional capacity and practice.