



Network of Networks on Impact Evaluation

■ DAC Evaluation Network ■ Evaluation Cooperation Group ■ International Organization for Cooperation in Evaluation ■ UN Evaluation Group

***‘You Can Get It If You Really Want’*: Impact
Evaluation Experience of the Office of Evaluation
and Oversight of the Inter-American Development
Bank**

Inder Jit Ruprah

NONIE WORKING PAPER NO. 3

January 2008

What is NONIE?

Nonie is a network of networks for impact evaluation comprised of the DAC Evaluation Network, The United Nations Evaluation Group (UNEG), the Evaluation Cooperation Group (ECG), and a fourth network drawn from the development evaluation associations (AfrEA, IOCE, IDEAS, ReLAC, and IPEN). Its purpose is to foster a program of impact evaluation activities based on a common understanding of the meaning of impact evaluation and approaches to conducting impact evaluation.

To this end a task team has been constituted and tasked with the following activities:

1. Preparation of impact evaluation guidelines
2. Agreeing collaborative arrangements for undertaking impact evaluation, leading to initiation of the program
3. Developing a platform of resources to support impact evaluation by member organizations

NONIE Working Papers

NONIE working papers present conceptual papers and impact evaluation findings. They may have been published elsewhere, e.g. as government or agency reports, but are included in the NONIE series to increase dissemination. Feedback on papers via the NONIE website is welcome.

***'You Can Get It If You Really Want'*¹: Impact Evaluation Experience of the Office of Evaluation and Oversight of the Inter-American Development Bank**

Inder Jit Ruprah²

Abstract

This paper's assessment of the Inter-American Development Bank's Office of Evaluation and Oversight's experience with impact evaluation offers lessons for best-practice methodology, including for studies faced with time and budget constraints. We point out the difficulty in mainstreaming this methodology in a multi- or bi-lateral lender. However, given its didactic nature, this assessment can be instructive to the development community, as we present solutions to rigorously evaluating programs that have not been designed with such an evaluation in mind.

I. Introduction

The international development community has been put on notice. The Center of Global Development asserts, 'For decades development agencies have disbursed billions of dollars ... Yet the shocking fact is that we have relatively little knowledge about the net impact of most of these programs.' (Savedoff and Levine 2006 and CGD 2006) The criticism is accompanied by a proposed minimum standard of knowledge: 'To determine what works ... It is necessary to collect data to estimate what would have happened without the program ... [only thus] ... It is possible to measure the impact that can be attributed to the specific program.' The criticism also contained a note of despair; and it called for an independent evaluation entity to ensure rigour in the evaluation of development programs.

This paper re-looks at the veracity of the assertion of the 'shocking fact' for the Inter-American Development Bank, a multi-lateral Bank that lends to Latin American and Caribbean countries, and whether the Bank's independent evaluation office, Office of Evaluation and Oversight (OVE), has made any difference. The paper also contributes to

the discussion regarding these criticisms of the international development community's lack of evaluative rigour. The paper mainly documents the experience of the OVE in carrying out impact evaluations, the asserted minimum standard of knowledge.

The story's relevance, however, is not limited to other evaluation offices of multi-lateral and bilateral organisations in the development community. The challenge faced by OVE, namely the *ex post* evaluations of projects that were neither designed for impact evaluation nor that collected outcome data, is probably the most common challenge faced by evaluators. In addition, OVE's experience adds to the growing evidence questioning the validity of the arguments against impact evaluations. The litany of arguments normally consists of: it is too difficult; it is too expensive; too few governments will agree; and there is no institutional mandate. Thus, the challenges faced by and the experience of OVE contribute to understanding the real world approaches to impact evaluations.

II. The Context

The Office of Evaluation and Oversight (OVE) was created in mid-1999 as part of the reform of the Bank's evaluation system. At that time OVE became independent of Bank Management, reporting solely to the Board of Executive Directors. In this redesign, the Board mandated OVE to: conduct Country Program Evaluations (CPE); conduct policy, strategy, thematic, and instrument evaluations; oversee the Bank's internal monitoring and evaluation system; oversee reviews of corporate strategy; provide normative guidance on evaluation issues; and contribute to evaluation capacity building in the region.

OVE did not have a mandate to evaluate individual operations. Only in 2003 did OVE receive a mandate to perform *ex-post* project evaluations. (IADB 2003) Thus, rather than being put on notice, the reason OVE took on this exercise was a change in Bank's policy.

The new policy mandated *ex post* project evaluations two to four years after a project closed. It said little to nothing about selection of what or how to evaluate or the minimum method standard that should be adopted. However, there was an assumption of stand-alone project evaluation and a method of before- completion-after reflexive type. The Bank would do the before-completion part and OVE would be relegated to the completion-after part.

However, the policy was based on false premises. First, the Bank does not routinely collect information for before-after or before-completion naïve reflexive evaluations.³ Generally, there is no full statement of development outcome intent at project approval. The Bank's system does not typically collect outcome information on on-going projects. The Bank's evaluations are almost void of statements on development outcomes upon closure (See Chart1). While it is necessary to collect data to estimate what would have happened without the program in order to determine what works, the Bank's evaluation system is not designed to do so; it does not typically even collect outcome information on beneficiaries.⁴

<Insert Chart 1>

Second, there is an assumption that outcomes can only be discerned years after a project has closed. However, other than lumpy investment loans, many, if not most, of the Bank's loans finance programs where development effects can be discerned a few years into the project. Third, the policy's focus was on the IADB projects. Often these are embedded in larger country programs. Thus, leaving aside the contribution to the design of a program, unless the benefit and the selection process of beneficiaries differ between the project and program, then the focus should be on the program not the project regarding development effectiveness. Finally, the policy emphasised the 'sustainability' of the program more in fiscal and institutional terms rather than the sustainability of the development effects.

Given this context, OVE decided to implement the *ex post* evaluation task within three principles: First, despite no institutional mandate, it decided to set impact methodology as

a minimum standard (Blundell and Costa 2002). Second, to conduct the impact evaluations using a theory-based approach (Fear 2007). Third, to adopt a purposeful rather than a random selection criterion of the programs to be evaluated, i.e., select similar projects within a thematic or meta-evaluation. OVE accepted that to determine ‘what works and what does not’ requires a quantitative approach, and within the quantitative approach, accepted the emerging consensus of a hierarchy of empirical evidence.⁵

The above principles were accompanied by decisions on how to implement the evaluation. The first issue was whether to carry out the evaluations in-house or to outsource them. The decision was to experiment with different modalities that covered all possibilities. The second issue was how to select consultants. The decision was to create a network of evaluators. The third issue was how to involve those evaluated, i.e. Bank staff and governments. The decision was to create a peer review group drawn from the Bank’s staff and another peer review group within the country.

III. Experience

In this section, we narrate OVE’s experience in carrying out impact evaluations. The success is judged with respect to numerous benchmarks: rigorous method standard, full implementation of the theory-based approach, meta-evaluations, the cost of the evaluations, the organisation of the task, and advocacy of impact techniques as a minimum standard.

Rigorous Method

If the standard of success is the use of counterfactuals to determine the impact of programs, then OVE has been successful. The Office has so far used the following impact techniques. Of the twenty-seven processed evaluations (i.e. publicly available), the techniques used have been, in order of importance: double difference with propensity score method (11), single difference with propensity score method (8), regression-instrument variable (5), and discontinuity regression method (1).⁶ Sometimes, for sensitivity or robustness reasons, more than one method in a given evaluation was used.

Often, naïve (i.e. before-after comparison of beneficiaries) or pipeline (i.e. comparison group composed of applicants to a program who have not yet received the program's benefit) techniques are included in OVE's impact evaluations.

In fact, the signature feature of OVE's *ex post* program evaluations is that they consist of routine comparisons between naïve (before-after or pipeline) and impact calculations. The reason for the comparison is essentially to advocate to the Bank that its task is not to fully implement its existing system based on an *ex post* comparison with a baseline but no comparison group, but rather to move towards a system that routinely involves impact evaluations. In Chart 2, the naïve and impact evaluations of a Social Investment Fund in Panama are shown using the change in poverty as the outcome. The naïve before-after calculation shows that poverty rose amongst the beneficiaries. The program was a failure. The impact calculation shows that the program's impact is a reduction in poverty. The program was successful. The example illustrates the 'you do not necessarily get what you see' reason for impact evaluations and that impact calculations are not always less than naïve ones.

<Insert Chart 2>

A priori, OVE expected to frequently use the regression discontinuity technique (Imbens and Lemieux 2007). High expectations were based on the assumptions that many programs had budget limits relative to the targeted population and the program's beneficiary selection process was based on ranking of applicants. However, *de facto* OVE has found it difficult to obtain the rankings and was therefore unable to use this technique. Perhaps the problem of non-availability is due to the continuing confusion between audits and evaluations. The only example is an evaluation of a Chilean Government Research Fund. The outcomes used were number and quality of publications. The impact calculations reveal that the program had no significant effect on outcomes. Chart 3 shows that the method is possible even when the accepted/non-accepted classification of applications to a program do not strictly follow the published ranking criteria of the program. In this case the method is fuzzy discontinuity. However,

the argument that even fuzzy data can be used does not reduce the fear that an evaluator is really an auditor.

<Insert Chart 3>

In contrast, OVE did not expect to be able to estimate an impact effect based on experimental data which, being *a priori*, is the ideal setting to perform unbiased impact evaluations.⁷ However, in the labour training thematic review, two random evaluations were feasible. One was the result of a well thought out evaluation design (Dominican Republic) and the other was from a natural experiment, in which a valid control group was *de facto* created due to an administrative cluster (Panama). Chart 4 shows the impact evaluation of the labour training program in the Dominican Republic which used random assignment. It shows that the program was successful for employability, income, and access to health insurance.

<Insert Chart 4>

The above example also shows that impact evaluations are often limited to answering whether there was a significant impact on the outcomes of interest. This is also the most common approach of OVE. However, policy concern also includes the issues of whether more budgetary outlay per capita increases the benefit, the dosage dimension of a program, and whether a multi-treatment has a greater impact than single-treatment. Chart 5 shows the impact calculations for Chile's government regional fund, the National Fund for Regional Development (FNDR). The transfers are mostly specific-purpose input based conditional, non-matching transfers. Chart 5 shows the different impacts of increased per capita transfers; there is no increase in poverty reduction above twelve times the base expenditure. The impact of transfers increases for diversified transfers (no one type of transfer is greater than 20 per cent of total transfers) vs. concentrated transfers (one type of transfer is 50 per cent or higher, in this case, for education) where the outcome is school attendance.

<Insert Chart 5>

Theory-based

If the benchmark for success is the systematic testing of all the links – the assumptions – in the causality chain of a given program, then OVE’s success has been partial. This partial success is due to budget restrictions and because it was often impossible to retrofit the required information.

A theory-based approach was adopted because it often gives plausibility to the impact findings. Theory or program based approaches map out the channels through which the activities, inputs, and outputs are expected to result in the expected outcomes. It also allows for the identification of unintended effects. Such mapping helps to identify key assumptions whose empirical validity could be tested for, allows an integration of contextual analysis including process evaluation that could account for the same program design having different performances, and possibly allows for the distinction between implementation failure and design failure. Not all these possible advantages have been fully exploited by OVE.

However, a distinction is often made between process evaluations and outcome evaluations, where impact evaluation is assumed to be only useful for determining outcomes. To the contrary, the impact technique can be used to evaluate process. For example, community participation is often asserted to have high dividends in terms of outcomes relative to non-community participation program delivery systems. Often, satisfaction surveys are taken as sufficient method to determine the success of a program. Chart 6 shows the impacts of community participation on the efficacy of a Social Investment Fund on school attendance and grade repetition as well as community satisfaction. The evaluation shows that if ‘dividend’ is taken to mean perceptions, i.e. community satisfaction, then the assertion is correct. If dividend is taken to mean an increase in outcomes, then it is incorrect- the impacts are statistically zero.

<insert Chart 6>

Impact techniques can also be used to check for the validity of key design features of a program. In Latin America many governments' social housing programs are based on the ABC (Spanish acronym for savings-grant-mortgage) design. High delinquency rates of publicly provided mortgages are often interpreted to be an example of intrinsic moral hazard of public provision. An interpretation often based on a probit regression with a dummy for the provider. The moral hazard interpretation leads to a call to change the provider from public to private. However, by using propensity score matching to obtain a valid comparison group (i.e. borrowers with similar relevant characteristics) and estimating the regression, the provider becomes irrelevant. The problem is incapacity to pay, hence redesign calls for the elimination of the mortgage component and a corresponding increase in the grant component. Chart 7 shows the marginal impact of mortgages provided by the public entity versus a private one. The marginal effect of public provision is a statistically significant increase in the probability of delinquency. As the right hand side of Chart 7 shows, the regression is based on very dissimilar households. Using the matched data, for the support group composed of similar households that received either a private or a public mortgage, the marginal effect of the provider becomes statistically zero.

<Insert Chart 7>

Meta-evaluations

If the standard of success is the systematic evaluation of similar programs across time and space then OVE has been relatively successful. The thematic approach, i.e. simultaneously evaluating similar programs, was adopted under the assumption that using a similar methodology, similar control variables, and a common set of outcomes would lend greater credibility to the evaluative findings of a given type of a program.

The first round of met-evaluations included: Youth Labour Training Programs; Science and Technology; and Rural Roads. The second round, which is in the advanced production stage, includes projects drawn from the following themes: Agricultural Technology Uptake, Social Investment Funds, and Early Childhood Development

programs. A third round, in early production stage, includes Citizen Security, Animal and Plant Health Systems, and Housing Programs.⁸

An example of a thematic evaluation is given in Table 1. A literature review of the impacts of active labour market programs in general and job training programs in particular, find modest results in OECD countries. There are little to no evaluations of these programs in Latin America. OVE analysed the experiences, applied the most robust methodology for each country, and then repeated the analysis with the same estimation technique in all the countries. The analysis concluded that there are significant impacts for particular groups, such as women and in some cases the youngest participants. In general, the impacts are larger for the quality of employment (i.e. formality) than for the gross employment rate.

<insert Table 1>

However, what theory rarely illuminates is the dynamic path of the benefits of a given intervention. The best that can be obtained is an unambiguous statement of steady state effects. Thus the timing of an impact evaluation may matter. Chart 8 shows the impacts on income and consumption of the Rural Road Rehabilitation program in Peru. It not only shows a different impact from motorised compared to non- motorised rural road rehabilitation, but also shows differing changes of those effects over time.

<insert Chart 8>

For example, in terms of sustainability of benefits, the evaluation of the job training program in the Dominican Republic illustrates the importance of continuous follow up. Chart 9 shows the impact of labour training on a given cohort over time. The short term

results (ten months after training) suggested limited impacts, however after more passage of time, positive impacts were detected- declining after a certain point, however.

<Insert Chart 9>

Costs

If the benchmark of success is judged by obtaining impact evaluations ‘on the cheap’ then OVE has been very successful. The high costs of impact evaluations are often invoked to explain away the lack of impact evaluations. The IBRD reports a cost of US\$300,000 to US\$500,000 per project adding to the fear of adopting an impact standard for evaluation (White 2006). OVE’s evaluations cost (staff time, travel costs, and consultants) is much less, averaging about US\$43,000.⁹

The lower financial costs follow from: First, selection bias, i.e. selecting themes or projects where there is a high *a priori* probability of finding existing data. OVE keeps costs down by not generally incurring primary data collection. Second, costs are reduced due to economies of scale obtained by evaluating a number of similar interventions simultaneously. Third, costs are less due to exploiting local expertise by using local consultants through a specially created network of evaluators, EVALNET. Local consultants have *a priori* knowledge of context, actors, program etc., which bypasses upfront learning costs and they usually charge less than similar evaluators from developed countries as there is reduced travel and interview costs. Most importantly, the network can be used to determine where the required data is available.

However, there are quality costs to this approach. The method adopted in the evaluations was due to the data available for the evaluation- not the other way round.¹⁰ Using existing secondary data has all the problems of the ‘tail wagging the dog.’ First, it implies an extremely high drop out rate of about 65 per cent. Second, not all desirable outcomes, intended or unintended, can be measured. Third, it is not always possible to determine the impact of a common set of outcomes using a common set of control variables and the same estimation technique across similar projects, which is the objective of a meta-

evaluation. This reduces comparability across evaluations. Fourth, it implies cutting corners- not necessarily by accepting a lower level of statistical precision, but by not being able to determine impacts at a lower level of disaggregation i.e. differentiating impacts by the different groups in the population being studied, and by not being able to evaluate all components of a given project.

Organisation

If the benchmark of success is judged by the determination of an ideal organisational structure that underlies the impact evaluation, then OVE has been unsuccessful. The first organisational dimension examined was which modality (in-house, outsourced, or in-between) was better. Of the 27 processed evaluations, 63 per cent were completely outsourced and 26 per cent were completely in-house. However, of the evaluations in progress most are mixed with the impact exercise being in-house but the context and program description is outsourced. The second organisational dimension was selection of consultants. The creation of a network of evaluators has been extremely useful. About 535 evaluators are fully registered, of which seventy per cent are Spanish speakers. The network has been very useful in searching both for in-country experts and for the existence of the necessary data for an impact evaluation.¹¹ The third dimension was mechanisms to involve the evaluated. The peer modality adopted by the Office has been unsuccessful both generally at obtaining input at the evaluation proposal stage, and at systematically entering into the Bank's design-evaluation-redesign of program cycle. The few examples of success are due to idiosyncratic reasons, i.e. individuals despite institutional resistance.

Advocacy

If the standard of success is the achievement of a systemic mainstreaming of impact evaluation within IDB, where impact evaluations are used routinely in the design and redesign of operations, then OVE has been unsuccessful.

This failure cannot be due to costs. The IADB approves annually about US\$6.5 billion, it annually disburses about US\$5 billion, its' annual research budget is US\$36 million, of the existing portfolio of loans there are US\$65 million nominally allocated (as part of a loan or in an associated technical grant) to evaluation. It cannot be due to staff without the required skills. Given the competitive salaries it pays, it could easily hire the expertise. It cannot be due to the lack of an institutional mandate. The absence of a mandate is self- imposed.

Failure can perhaps be attributed to two reasons. On an individual professional level it could be argued that it pays to be ignorant (Pritchett 2002). Publicly available impact findings are neither of interest to the IDB's operation officers responsible for the loan, nor for the staff of the program's executing agency that represents the government which contracted the loan. A finding of a zero impact plus a high present value of the debt incurred would be politically inconvenient for both parties.¹² At the institutional level, there are clearly tradeoffs between the different uses of evaluation:

They are simultaneously used as an instrument of transparency and control, accountability, legitimization and institutional learning. With respect to the legitimization function, evaluation can be thought of as marketing device to prove the aid organization's successful work to the general public However ... the legitimization function seems to be dominating. Transparency and legitimization are clearly conflicting objectives in all cases in which actual development outcomes are not fully satisfactory (Michaelowa and Borrmann 2005).

An external evaluation office would by its very nature be viewed as one whose role is accountability, and would by institutional design be outside the design-implementation – experience- redesign learning cycle.

The failure is also due to OVE. There cannot be advocacy if there is no effective dissemination policy. Dissemination through documents, seminars, conferences etc. have

been crowded out in order to meet the increasing targets for the number of projects evaluated.

IV. Conclusions

The asserted ‘shocking fact’ of ignorance of development effects is correct for IDB. It is not correct if OVE is taken into consideration. OVE’s experience suggests that ‘You can get it if you really want.’

OVE’s experience shows that the arguments that impact evaluations are too difficult, too expensive, too few governments will support them, and too far from feasibility without an institutional mandate do not hold. They are not too difficult, like everything they just require the appropriate skills. They are not too expensive. They are not opposed by most governments, once they know that it does not involve budget costs. They can be carried out independently of an institutional mandate.

Thus, if the benchmark of success is the production of a large number of rigorous evaluations then OVE’s story is one of exceptional success. This benchmark is inappropriate, however. Success should be measured by the degree that impact evaluations are adopted as the norm in the institution. This has not occurred. The demonstration effect is non-existent. Success could also be judged by the creation of an effective virtuous cycle of institutional learning, whereby independent evaluation leads to lessons identification and lesson utilisation by the institution leads to improved operational work that in turn leads to improvement in lives. This has not entirely materialised. The few examples of success are due to idiosyncratic factors not institutional ones.

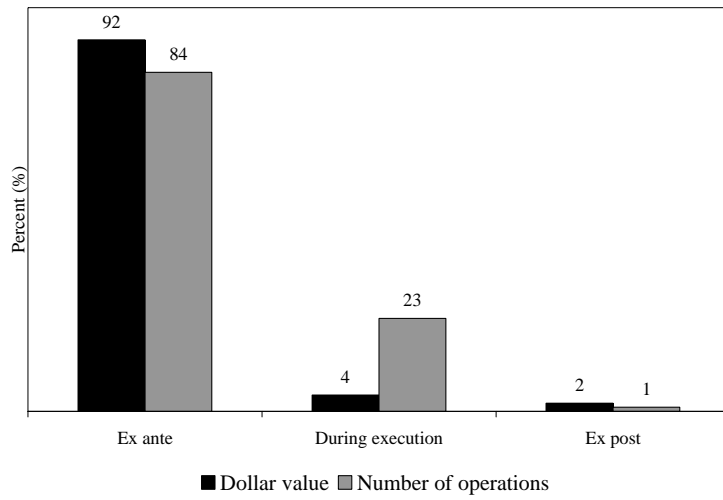
Thus OVE’s experience bodes ill for the proposed independent international evaluation entity. The challenge is not the feasibility of impact evaluations at the retail level; OVE’s experience reveals that is entirely feasible. The real challenge is to succeed in convincing the actors in the international development community to measure the impact of their programs and in doing so to obtain the scale needed for an effective virtuous cycle of

improving lives through evaluation. After four years, OVE has been unable to convince its own institution of the virtues of impact evaluation.

Annex 1: OVE Impact Evaluations

- Alzuá M. and Brassiolo P. The Impact of Training Policies in Argentina: An Evaluation of Proyecto Joven WP-15-06. Washington DC: OVE, 2005.
- Binelli C. and Maffioli A. Evaluating the Effectiveness of Public Support to Private R&D: Evidence from Argentina. WP-11-06. Washington DC: OVE, 2005.
- Chudnovsky D., Andrés López, Martín Rossi and Diego Ubfal. Evaluating a Program of Public Funding of Private Innovation Activities. An Econometric Study of FONTAR in Argentina. WP-16-06. Washington DC: OVE, 2005.
- Chudnovsky D., Andrés López, Martín Rossi and Diego Ubfal. Evaluating a Program of Public Funding of Scientific Activity. A Case Study in Argentina WP-12-06
- Davis, B., Sudhanshu Handa, Marta Ruiz, Marco Stampini and Paul Winters. An Impact Evaluation of Agricultural Subsidies on Human Capital Development and Poverty Reduction: Evidence from Rural Mexico. WP-03-05. Washington DC: OVE, 2005.
- Delajara M., Samuel Freije, and Soloaga I. An Evaluation of Training for the Unemployed in Mexico. WP-09-06. Washington DC: OVE, 2005.
- De Negri J.A. , Lemos M., and De Negri F. The Impact of University Enterprise Incentive Program on the Performance and Technological Efforts of Brazilian Industrial Firms. WP-13-06. Washington DC: OVE, 2005.
- De Negri J.A. , Lemos M., and De Negri F. Impact of P&D Incentive Program on the Performance and Technological Efforts of Brazilian Industrial Firms. WP-14-06. Washington DC: OVE, 2005.
- Díaz, J.J. and Handa, S. An Assessment of Propensity Score Matching as a Non Experimental Impact Estimator: Evidence from Mexico's PROGRESA. Program WP-04-05. Washington DC: OVE, 2005.
- Díaz J.J. and Jaramillo M. An Evaluation of the Peruvian "Youth Labor Training Program" – PROJOVEN. WP-10-06. Washington DC: OVE, 2005.
- Galdo, V. and Briceño B. An Impact of a Potable Water and Sewerage Expansion in Quito: Is Water Enough? WP-01-05. Washington DC: OVE, 2005.
- Heinrich C. Demand and Supply-Side Determinants of Conditional Cash Transfer Program Effectiveness: Improving the First-Generation Programs. WP-05-05. Washington: OVE, 2005.
- Heinrich C. and López Y. Does Community Participation Produce Dividends in Social Investment Fund Projects? WP-01-07. Washington DC: OVE, 2005.
- Heinrich C. and Cabrol M. An Impact Evaluation of the National Student Scholarship Program in Argentina. WP-06-05. Washington DC: OVE, 2005.
- Marcano, L. Una Evaluación de Impacto del Programa de Fondo de Inversión Social de Panamá. WP-02-05. Washington DC: OVE, 2005.
- Torero M. and Field E. An Impact Evaluation of Land Titles on Rural Households in Peru. WP-07-05
- Ureta, B. Cocchi H. and Solis D. Output Diversification Among Small-Scale Hillside Farmers in El Salvador. WP-17-06. Washington DC: OVE, 2005.
- Ureta, B. Cocchi H. and Solis D. Adoption of Soil Conservation Technologies in El Salvador: A Cross-Section and Over-Time Analysis. WP-19-06. Washington DC: OVE, 2005.
- Ureta B. and Cocchi H. On-Site Costs and Benefits of Soil Conservation Among Hillside Farmers in El Salvador. WP-19-06. Washington DC: OVE, 2005.
- Ruprah I. and Marcano L. A Meta-Impact Evaluation of Social Housing Programs: The Chilean Case. WP-02-07. Washington DC: OVE, 2005.
- Soares F. and Soares Y. The Socio-Economic Impact of Favela-Bairro: What do the Data Say?. WP-08-05. Washington DC: OVE, 2005.

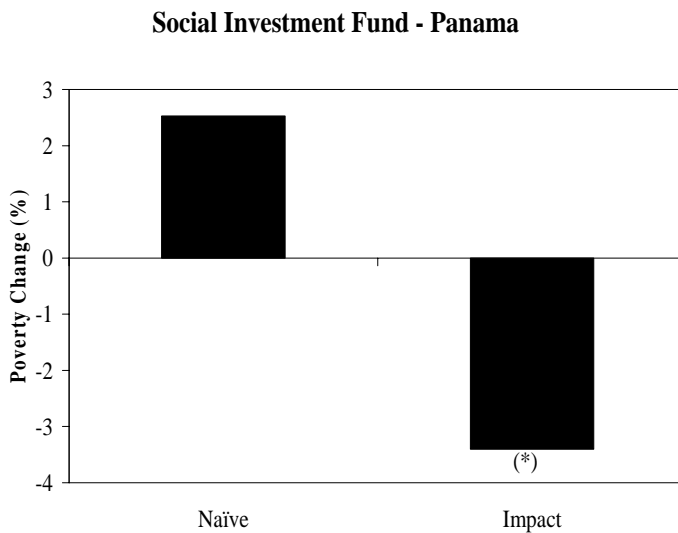
Chart 1. Information on Outcomes of IADB Operations.



- Ex ante evaluability (83 investment projects approved in 2005). 84% had information (baseline and target) on one outcome; only 6 had full information set.
- Evaluability of on-going projects (420 active projects in 2006). 23% reported that they were collecting information on at least one outcome indicator.
- Ex post evaluability (38 projects that closed in 2005.) 1% had data for a naïve evaluation.

Source: OVEDA

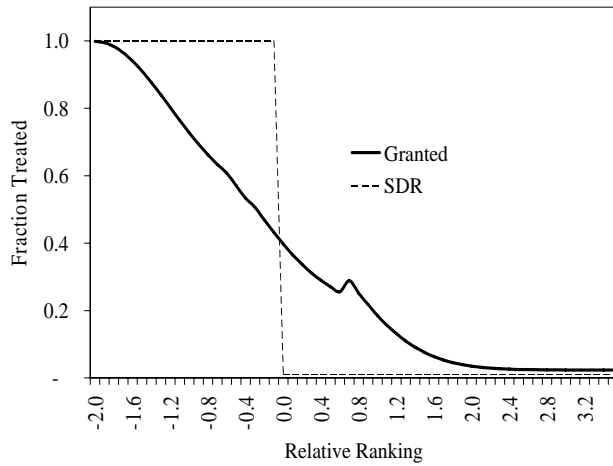
Chart 2: Naïve Vs Impact



- Profile: Social Investment Fund Panama; Basic Infrastructure to poor communities
- Data: Distribution of benefits by municipalities from administrative data; Baseline and results of outcome indicators from households surveys 1994-2001
- Technique: Treatment and comparison group using PSM in double difference. The sample included 75 municipalities. Potential to work with a sample of more than 250 smaller geographic units but household survey was not representative at that level
- Results: Naïve evaluation: the program failed. Impact evaluation: the program succeeded
- Cost. & Modality: US\$22,944 in-house

Source: Marcano (2005)

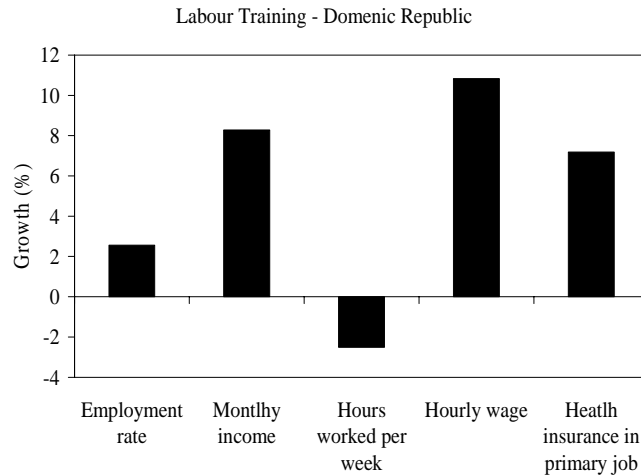
Chart 3: Fuzzy Discontinuity



- Profile: Science and Technology – Chile. Financing for research projects
- Data: Administrative data of all applicants. Ratings of all applicants and identification of accepted and rejected applicants and publications recorded in the ISI – SCI
- Technique: Discontinuity regression design. The selection process drawn by a “threshold” quality value that separates beneficiaries from non-beneficiaries
- Results: Unsuccessful. FONDECYT has no significant positive impact on the scientific production of the financed projects.
- Costs & modality: US\$25,000 & joint

Source: Benavente, Crespi and Maffioli (2007)

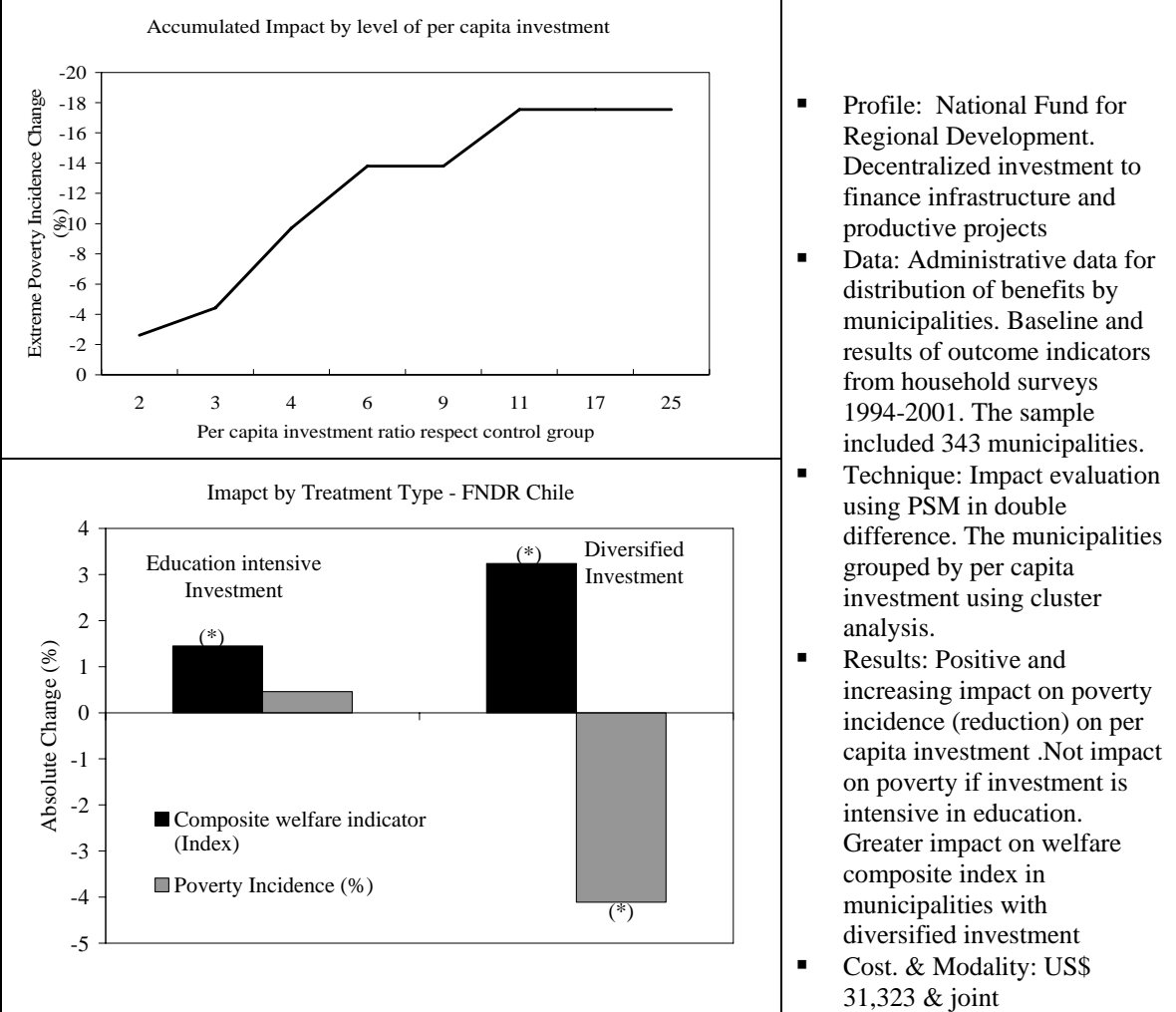
Chart 4: Labour Training and Random Treatment Impact Estimation



- Profile: Labour Training program, in the Dominican Republic
- Data: Simple randomization including a follow-up survey done at 10-14 months after graduation from training. 786 treated and 563 controls. Baseline has universe, follow up was a stratified random sample (size determined by standard formulas)
- Technique: Estimated Average Intention-to-treat on treated by simple diff of means, verified with weighted diff and regression analysis (no Difference in difference because of a faulty baseline).
- Results: Employability, income and health insurance access increased.
- Cost & modality: US\$31,000 & joint

Source: P. Ibararan et al (2006)

Chart 5: Dosage and Multi-treatment Impacts of a Regional Transfer Fund



- Profile: National Fund for Regional Development. Decentralized investment to finance infrastructure and productive projects
- Data: Administrative data for distribution of benefits by municipalities. Baseline and results of outcome indicators from household surveys 1994-2001. The sample included 343 municipalities.
- Technique: Impact evaluation using PSM in double difference. The municipalities grouped by per capita investment using cluster analysis.
- Results: Positive and increasing impact on poverty incidence (reduction) on per capita investment. Not impact on poverty if investment is intensive in education. Greater impact on welfare composite index in municipalities with diversified investment
- Cost. & Modality: US\$ 31,323 & joint

Source: Ruprah and Marcano (2007)

Chart 6: The Impact of Community Participation

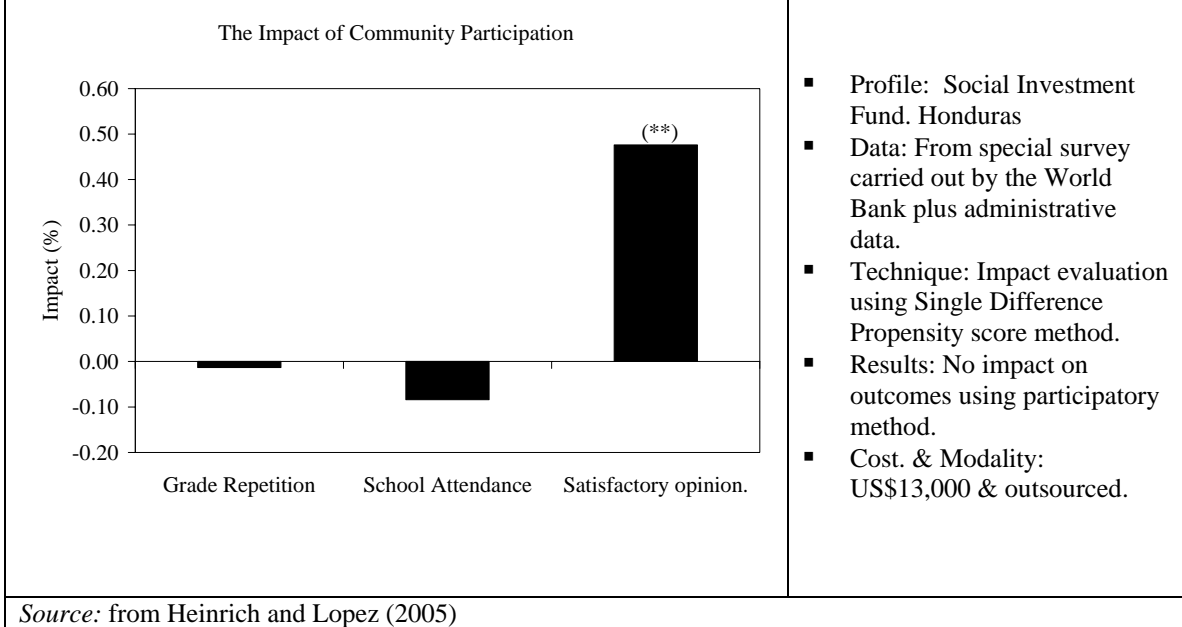
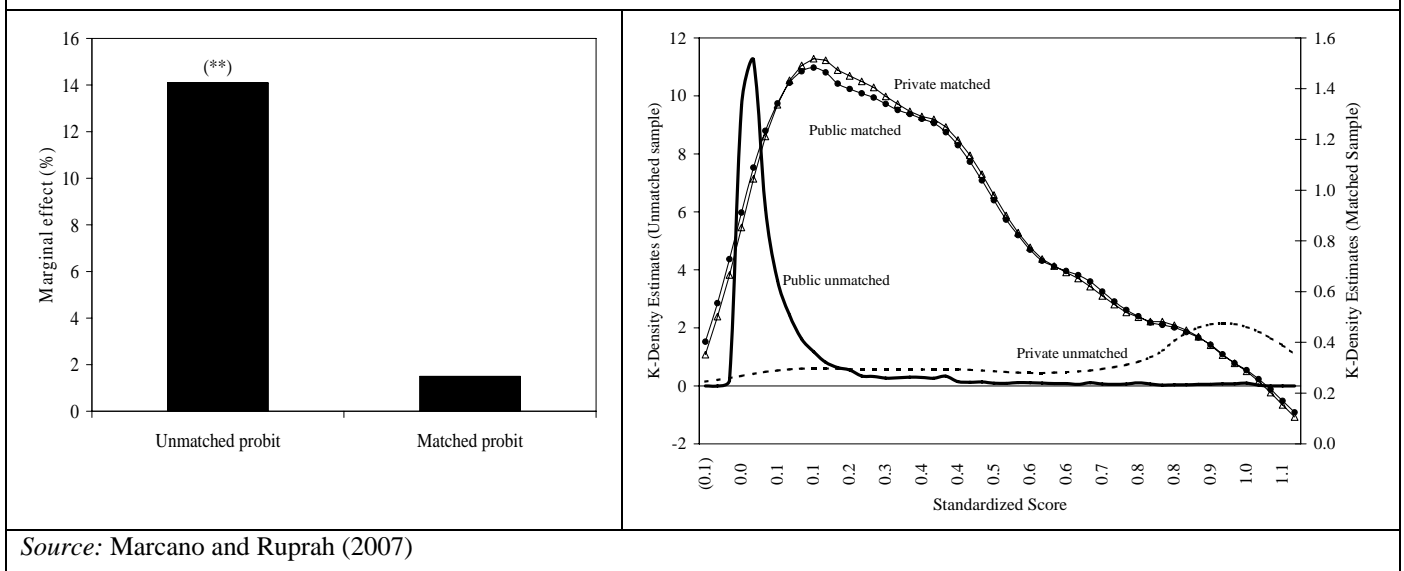
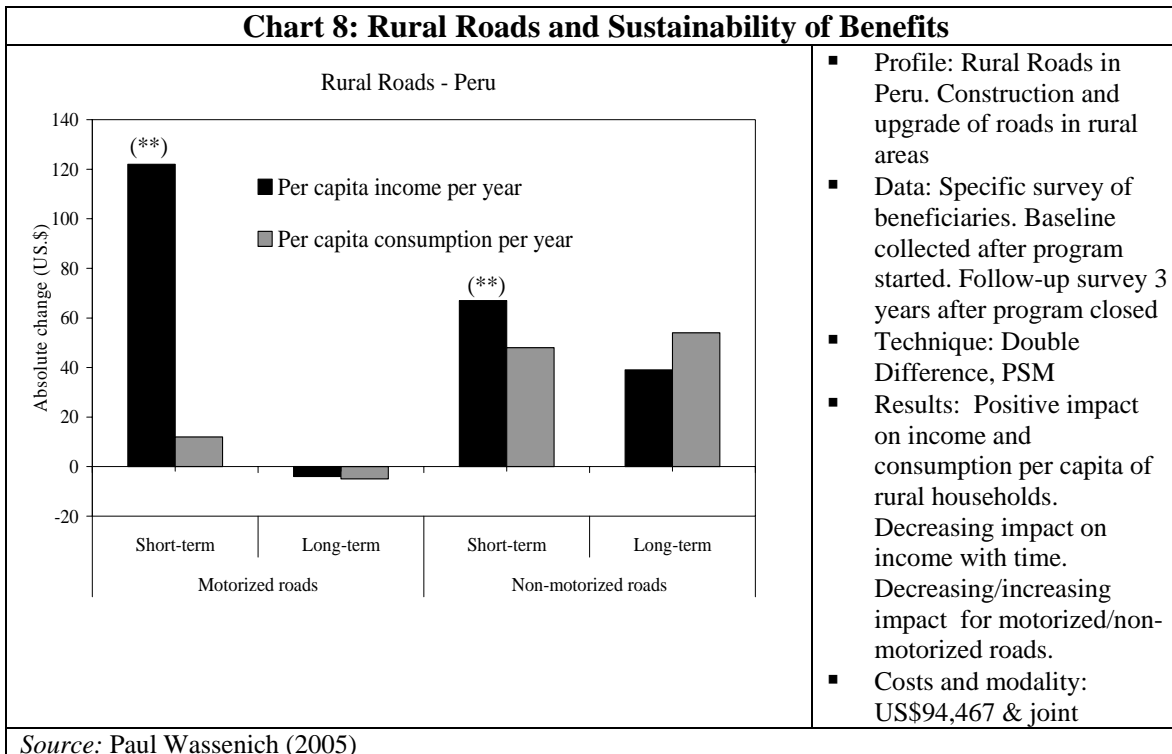


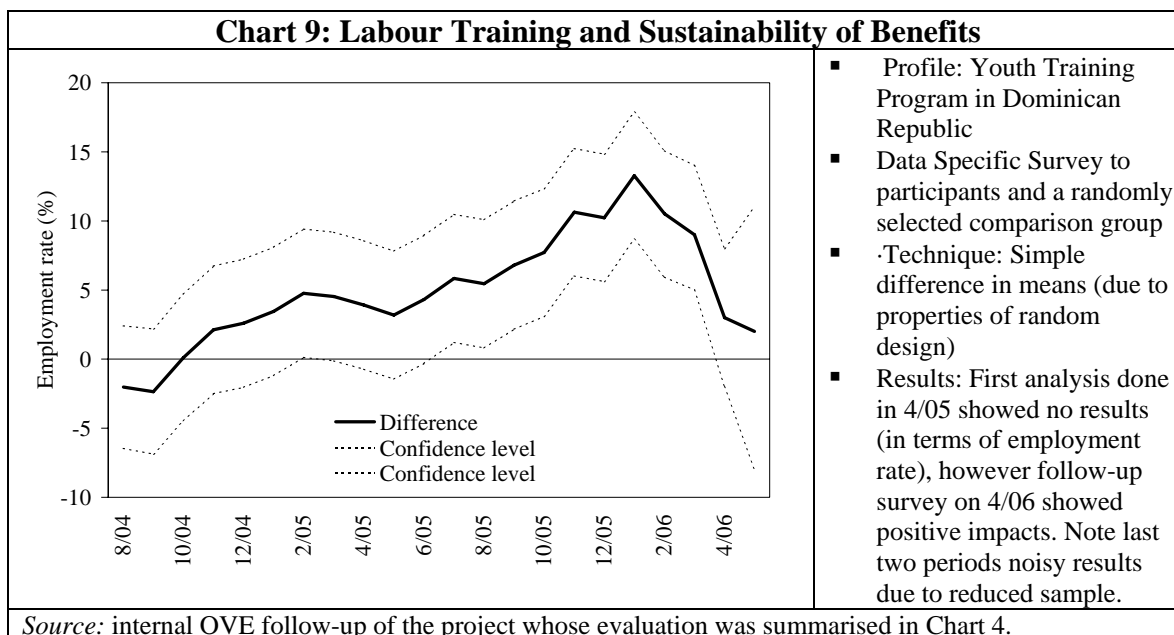
Chart 7: Mortgage Delinquency Rates: Moral Hazard or Incapacity to Pay.



	Methodology & Data	Employment Rate	Formality	Wages
Argentina	quasi-experimental, four rounds, primary data	10-30% for youngest (<21)	0% - 3%, 6% - 9% for young males	not significant
Dominican Republic	experimental, one round, primary data	None, higher (5-6%) but not significant in the East & Sto Dom	Health insurance 9% higher for men (43% vs 34%)	17% (sign. at 10%), larger for males under 19
Mexico	quasi-experimental, six rounds, primary data	No clear pattern for general employment	10-20% for salaried workers, 0-20% for self-employed, higher since 2002	no consistent patterns, at best small and mostly not significant
Panama	natural experiment, one round, primary data	0% - 5%, 10-12% for women and in Panama	overall not significant, probably higher outside Panama City	overall negligible, large for women (38%) and in Panama 25%
Peru	quasi-experimental, five rounds, primary data	13% (much higher for women --20% than for men --negligible)	11% (14% women, 5% men)	not significant

Source: Ibarra and Rosas 2006





References

Blundell R. and M. Costa (2002) *Alternative Approaches to Evaluation in Empirical Microeconomics* CEMMAP Working Paper CWP1002, London: UCL

Center for Global Development (2006) ‘Will We Ever Learn? Improving the Lives through Impact Evaluations’ Washington, DC: Center for Global Development

Coryn, Chris L. S. (2007) “The Holy Trinity of Methodological Rigor: A Sceptical View”. *Journal of MultiDisciplinary Evaluation*, Vol.4, Number 7.

Davidson, E.J. (2007) “The RCTs-only Doctrine: Brakes on the Acquisition of Knowledge?”, *Journal of MultiDisciplinary Evaluation*.

Dufflo E., and M. Kramer (2005) ‘The Use of Randomization in the Evaluation of Development Effectiveness’ in G.K. Pitman, O. Feinstein, and G. Ingram (Eds.) *Evaluating Development Effectiveness*. IEG, World Bank, 2005

Fear, W.J. (2007) ‘Program Evaluation Theory: The Next Step Towards a synthesis of Logic Models and Organisational Theory’, *Journal of Multidisciplinary Evaluation* 4.7

IADB (2003) “Ex Post Policy of Operations”, GN-2254-5, Sep 2003

Ibarraran, P. and D. Rosas, ‘IDB’s Job Training Operations: Thematic Report of Impact Evaluations’, OVE, IADB, Oct.2006, unprocessed.

Imbens, G. and Lemieux, T. (2007) *Regression Discontinuity Designs: A Guide to Practice* NBER Technical Working Paper 337, Cambridge: National Bureau of Economic Research

Michaelowa K. and A. Borrmann (2005) *What determines Evaluation Outcomes? Evidence from Bi and Multilateral Development Cooperation* HWWA Discussion Paper 310, Hamburg: Hamburg Institute of International Economics

Pritchett, Lant (2002) 'It Pays to be Ignorant: a Simple Political Economy of Rigorous Program Evaluation', *Public Reform* 5.4: 251-209

Savedoff, William and Ruth Levine (2006) *Learning from Development: The case for International Council to Catalyse Independent Impact Evaluations of Social Sector Interventions* CGD Brief, Washington, DC: Center for Global Development

White, Howard (2006) "Impact Evaluation: The Experience of the Independent Evaluation Group of the World Bank", IEG, World Bank.

Notes

¹ The title of a reggae song by Desmond Dekker.

² Of the Office of Evaluation and Oversight of the IADB, I would like to thank for their input particularly Luis Marcano and Pablo Ibarra and also Allesandro Maffioli, Yuri Soares and Ana Santiago, all members of OVE who are involved in ex post impact evaluations.

³ For a summary of the first year's experience of OVE and an evaluation of the Bank's monitoring and Evaluation system see OVE's report: Ex Post Project Evaluation:2004 Annual Report. AE-112. August 2005.

⁴ A couple of disheartening examples that OVE has come across are the following. In one case data collected by the Bank that could have been used for an impact evaluation were thrown out. The reason offered by Bank staff was that it was contaminated as it contained identifiable beneficiaries and non beneficiaries of the program! Another example is OVE staff was welcomed by an executing agency of a Bank program, they were happy that someone had come to collect the boxes that were using valuable space. The boxes contained sequential surveys still in paper form for evaluating a watershed project. Years of water and rats, however, precluded their use.

⁵ This brief summary of the "principles" papers over a heated discussion of methodology within OVE and between OVE and the Bank. Within OVE the discussion ranged from taking photographs of before and after to that only random trails were acceptable. Many of the points discussed parallel the succinctly review of the debate on methodology standards by Coryn (2007)

⁶ All of OVE's evaluations are made public. In the case of individual program's ex post reports, perhaps to the Office's under-investment in dissemination, there is an increasing stock of unprocessed (reviewed, formatted, and put on the web) reports.

⁷ For this argument see Dufflo and Kramer (2005). For an agnostic view, see Davidson (2007).

⁸ Stand-alone program evaluations were previously studied for two of these themes. See for the housing case Ruprah and Marcano (2007) and, for Citizen security case I.J. Ruprah and Luis Marcano, "Safer Chile: an Impact Evaluation of Chile's Citizen Security Program", 2007, not processed. However, the idea that first a stand-alone evaluation, and then, armed with the experience, extend to other similar programs has not been typical.

⁹ Note, that the total cost of the ex post project evaluation exercise is higher than the sum of the costs of the individual program evaluations. There is a high attrition rate i.e. most of the projects selected for the meta-evaluations are abandoned as no data for an impact can be found.

¹⁰ Of the 27 processed evaluations 74% used existing surveys and only 18% used surveys commissioned by OVE.

¹¹ This has required detailed terms of references that are sent out through EVALNET in which the terms of reference asks for an evaluation with the options hence price of: (i) existing data; or (ii) new data, or (iii) combination of both.

¹² Any impact design and evaluation by the Bank is due to idiosyncratic factors such as professional interest of the operational officer or government. However, these are not part of the routine evaluation system.