'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 Ruprah[2]

1 Introduction

The international development community has been put on notice. The Center for 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; 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 is it] ... 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 programmes.

This article reconsiders the veracity of the assertion of the 'shocking fact' for the Inter-American Development Bank (IADB), a multilateral Bank that lends to Latin American and Caribbean countries, and whether the Bank's independent evaluation office, the Office of Evaluation and Oversight (OVE), has made any difference. The article also contributes to the discussion regarding these criticisms of the international development community's lack of evaluative rigour. The article 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 multilateral and bilateral organisations in the development community. The challenge faced by OVE, namely the *ex post* evaluations of projects that were not designed for impact evaluation and didn't collect 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.

2 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 Programme Evaluations (CPEs); 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).

Figure 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 six had full information set
- Evaluability of ongoing 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.

Thus, rather than being put on notice, the reason OVE took on this exercise was a change in Bank policy.

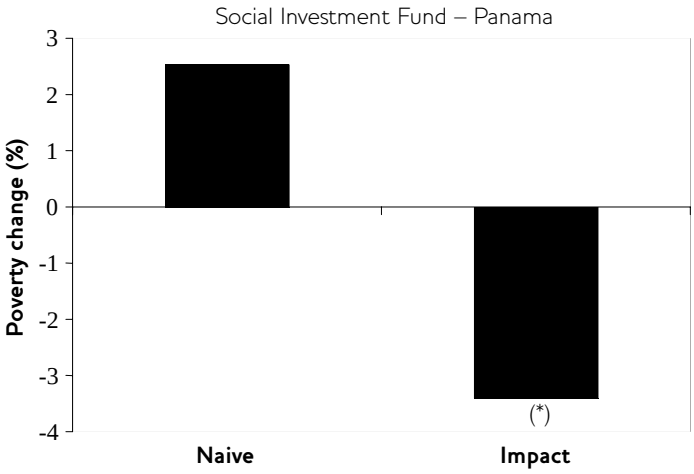
The new policy mandated *ex post* project evaluations two to four years after a project closed. It said little to nothing about selection of what or how to evaluate or the minimum method standard that should be adopted. However, there was an assumption of standalone project evaluation and a method of before-completion-after naïve 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 the necessary information on outcomes (neither baseline values nor their values near the time of project completion) for the 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 ongoing projects. The Bank's evaluations are almost void of statements on development outcomes upon closure (see Figure 1). While it is necessary to collect data to estimate what would have happened without the programme 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]

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 programmes in which 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 programmes. Thus, leaving aside the contribution to the design of a programme, unless the benefit and the selection process of beneficiaries differ between the project and programme, then the focus should be on the programme not the project regarding development effectiveness. Finally, the policy emphasised the 'sustainability' of the programme more in fiscal and institutional terms rather than in terms of the sustainability of the development effects.

Given this context, OVE decided to implement the *ex post* evaluation task on the basis of 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 programmes 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]

Figure 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 programme succeeded
- Cost: US$22,944
- Modality: in-house

Source Marcano (2005).

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.

3 Experience

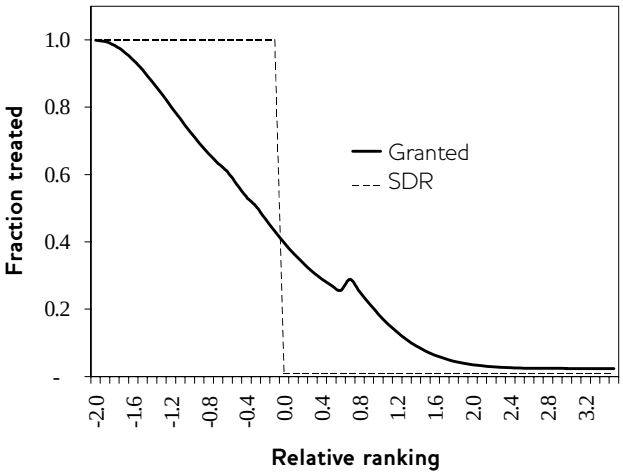
In this section, we narrate OVE's experience in carrying out impact evaluations. The success is judged with respect to numerous benchmarks: rigorous method standard, full implementation of the theory-based approach, meta-evaluations, the cost of the evaluations, the organisation of the task, and advocacy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

3.1 Rigorous method

If the standard of success is the use of counterfactuals to determine the impact of programmes, then OVE has been successful. The Office has so far used the following impact techniques. Of the 27 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 (7), 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 programme who have not yet received the programme's benefits) techniques are included in OVE's impact evaluations.

In fact, the signature feature of OVE's *ex post* programme 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 programme was a failure. The impact calculation shows that the programme's impact is a reduction in poverty. The programme was

Figure 3 Fuzzy discontinuity



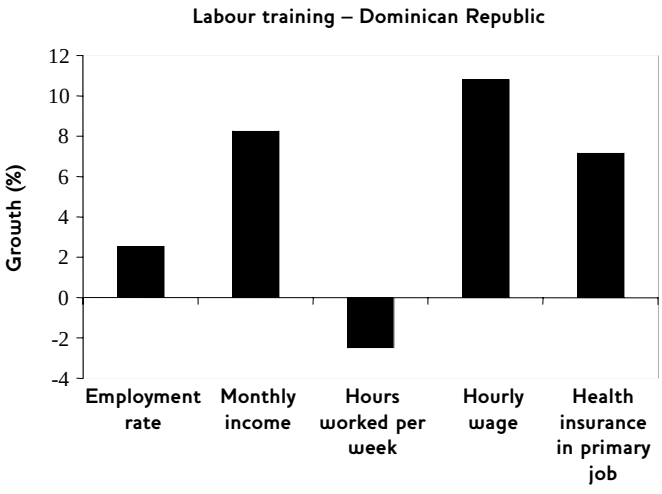
- Profile: Science and Technology – Chile. Financing for research projects
- Data: Administrative data of all applicants. Ratings of all applicants and identification of accepted and rejected applicants and publications recorded in the ISI – SCI
- Technique: Discontinuity regression design. The selection 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.
- Cost: US$25,000
- Modality: joint (staff and external consultants)

Source Benavente *et al.* (2007).

successful. The example illustrates the 'you do not necessarily get what you see' reason for impact evaluations and the fact that impact calculations are not always less than naïve ones.

A priori, OVE expected to frequently use the regression discontinuity technique (Imbens and Lemieux 2007). High expectations were based on the assumptions that many programmes had budget limits relative to the targeted population and the programme'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

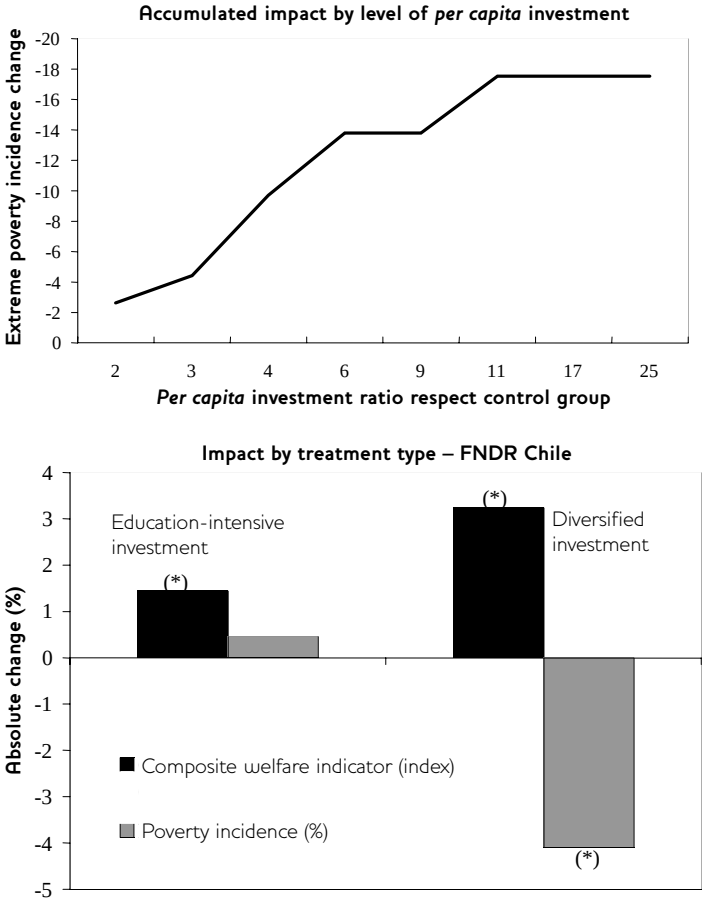
Figure 4 Labour training and random treatment impact estimation



- Profile: Labour Training programme, in the Dominican Republic
- Data: Simple randomisation 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 formulae)
- 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: US$31,000
- Modality: joint (staff and external consultants

Source Ibarraran and Rosas (2006).

Figure 5 Dosage and multi-treatment impacts of a regional transfer fund



- Profile: National Fund for Regional Development. Decentralised 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. No impact on poverty if investment is intensive in education. Greater impact on welfare composite index in municipalities with diversified investment
- Cost: US$ 31,323
- Modality: joint (staff and external consultants)

Source Ruprah and Marcano (2007).

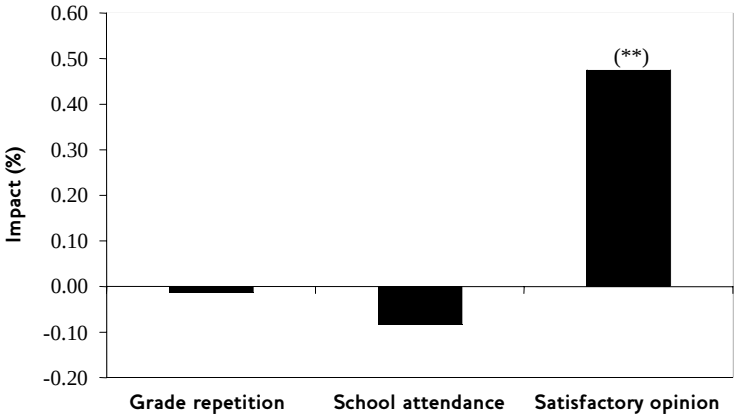
example is an evaluation of a Chilean Government Research Fund. The outcomes used were number and quality of publications. The impact calculations reveal that the programme had no significant effect on outcomes. Figure 3 shows that the method is possible even when the accepted/non-accepted classification of applications to a programme do not strictly follow the published ranking criteria of the programme. 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.

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 in which 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). Figure 4 shows the impact evaluation of the labour training programme in the Dominican Republic which used random assignment. It shows that the programme was successful for employability, income, and access to health insurance.

The above example also shows that impact evaluations are often limited to establishing 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

Figure 6 **The impact of community participation**



- Profile: Social Investment Fund, Honduras
- Data: From special survey carried out by the World Bank plus administrative data
- Technique: Impact evaluation using single difference propensity score method
- Results: No impact on outcomes using participatory method
- Cost: US$13,000
- Modality: outsourced

Source Heinrich and Lopez (2005).

issues of whether more budgetary outlay *per capita* increases the benefit, the dosage dimension of a programme, and whether a multi-treatment has a greater impact than single treatment. Figure 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. Figure 5 shows the different impacts of increased *per capita* transfers; there is no increase in poverty reduction above 12 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.

3.2 Theory-based

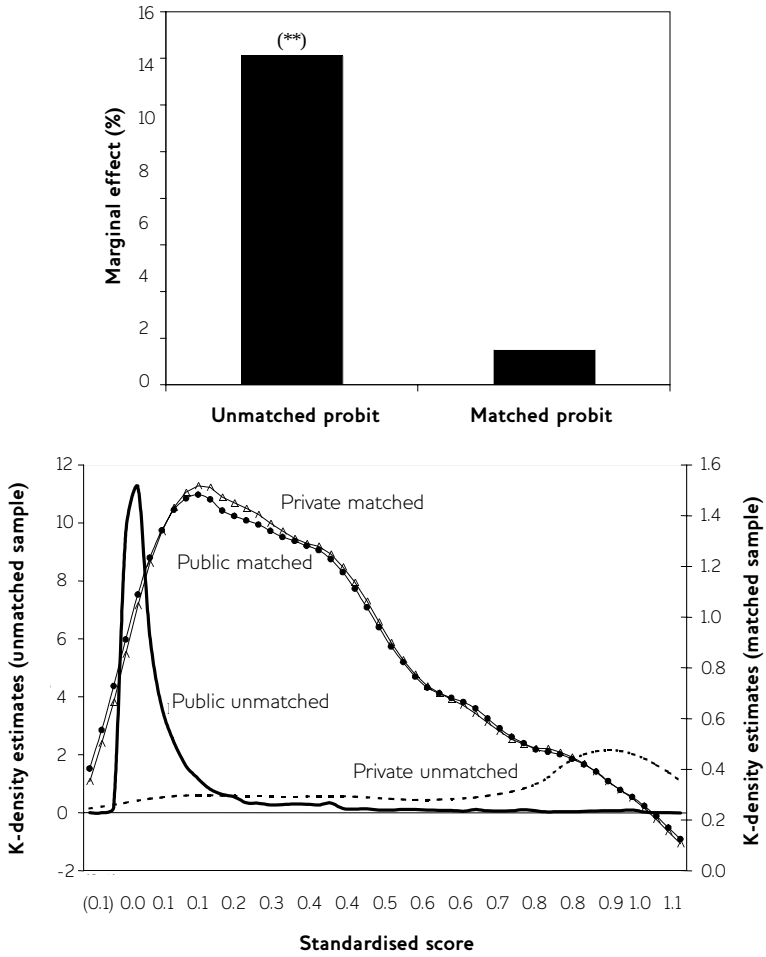
If the benchmark for success is the systematic testing of all the links – the assumptions – in the causality chain of a given programme, then OVE's success has been partial. This partial success is due to budget restrictions and because it was often impossible to retrofit the required information.

A theory-based approach was adopted because it often gives plausibility to the impact findings. Theory- or programme-based approaches map out the channels through which the activities, inputs, and outputs are expected to result in the expected outcomes. They also allow 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 programme design performing differently, 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. On 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 programme delivery systems. Often, satisfaction surveys are taken as sufficient to determine the success of a programme. Figure 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.

Impact techniques can also be used to check for the validity of key design features of a programme. In Latin America many government social housing programmes are based on the ABC (Spanish

Figure 7 Mortgage delinquency rates: moral hazard or incapacity to pay



Source Marcano and Ruprah (2007).

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. This is an interpretation often supported by 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. Figure 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 Figure 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.

3.3 Meta-evaluations

If the standard of success is the systematic evaluation of similar programmes across time and space then OVE has been relatively successful. The thematic

Table 1 **Summaries of individual labour training evaluations**

	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 & Santo Domingo	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 Ibarraran and Rosas (2006).

approach, i.e. simultaneously evaluating similar programmes, was adopted on 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 programme.

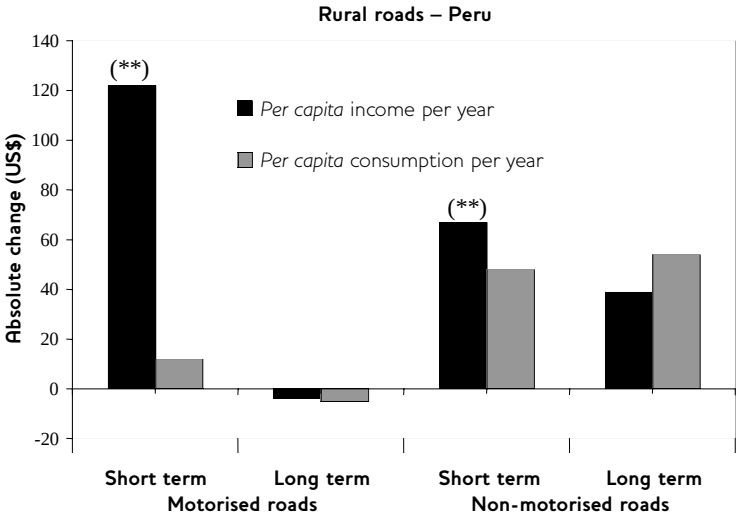
The first round of meta-evaluations included: youth labour training programmes; science and technology; and rural roads. The second round, which is in the advanced production stage, includes projects investigating agricultural technology uptake, social investment funds, and early childhood development programmes. A third round, in early production stage, includes citizen security, animal and plant health systems, and housing programmes.[8]

An example of a thematic evaluation is given in Table 1. A literature review of the impacts of active labour market programmes in general and job training programmes in particular, finds modest results in OECD countries. There are very few evaluations of these programmes 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.

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. Figure 8 shows the impacts on income and consumption of the Rural Road Rehabilitation programme in Peru. It not only shows a different impact from motorised as opposed to non-motorised rural road rehabilitation, but also shows differing changes of those effects over time.

For example, in terms of sustainability of benefits, the evaluation of the job training programme in the Dominican Republic illustrates the importance of continuous follow-up. Figure 9 shows the impact of

Figure 8 Rural roads and sustainability of benefits



- Profile: Rural Roads in Peru Construction and upgrade of roads in rural areas
- Data: Specific survey of beneficiaries. Baseline collected after programme started. Follow-up survey 3 years after programme closed.
- Technique: Double difference, PSM
- Results: Positive impact on income and *per capita* consumption of rural households. Decreasing impact on income with time Decreasing/increasing impact for motorised/non-motorised roads
- Cost: US$94,467
- Modality: joint (staff and external consultants)

Source Wassenich (2006).

labour training on a given cohort over time. The short-term results (ten months after training) suggested limited impacts; however after more time had elapsed, positive impacts were detected – albeit declining after a certain point.

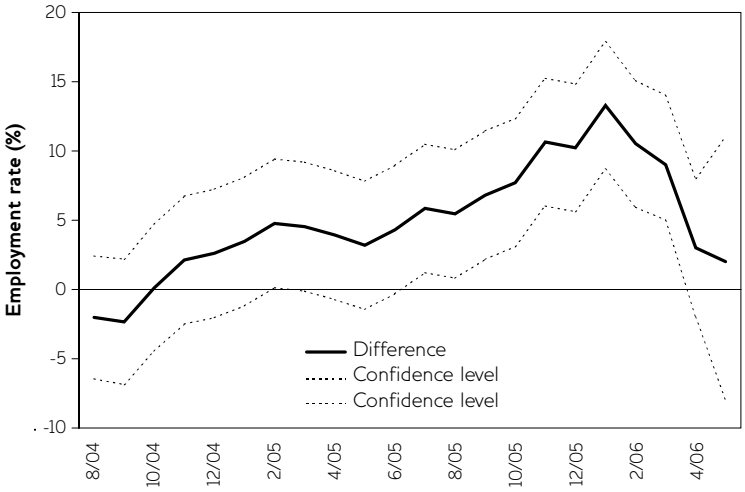
3.4 Costs

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 the lack of impact evaluations. The World Bank reports a cost of US$300,000–500,000 per project, adding to the fear of adopting an impact standard for evaluation (White 2006). The costs of OVE evaluations (staff time, travel costs, and consultants) are much lower, 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, programme etc., which bypasses upfront learning costs and they usually charge less than similar evaluators from developed countries as there are 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 dropout 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

Figure 9 Labour training and sustainability of benefits



- Profile: Youth Training Programme in Dominican Republic
- Data: specific survey of participants and a randomly selected comparison group
- Technique: Simple difference in means (due to properties of random design)
- Results: First analysis done in 4/05 showed no results (in terms of employment rate). However, follow-up survey on 4/06 showed positive impacts. Note last two periods noisy results due to reduced sample

Source Internal OVE follow-up of the project whose evaluation was summarised in Figure 4.

population being studied, and by not being able to evaluate all components of a given project.

3.5 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 programme description 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 whom 70 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 cycle of design-evaluation-redesign of programme. The few examples of success are due to idiosyncratic reasons, i.e. individuals despite institutional resistance.

3.6 Advocacy

If the standard of success is the achievement of a systemic mainstreaming of impact evaluation within IADB, 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 of interest to neither the IADB's operations officers responsible for the loan, nor for the staff of the programme's executive 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 trade-offs between the different uses of evaluation:

Figure 10 Photographs of before/completion of a slum upgrading project in Favela Barrios, Brazil

Source IADB. (The IADB maintains a database of photographs of most of its projects.)

> 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 a 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.

4 Conclusions

The asserted 'shocking fact' of ignorance of development effects is correct for IADB. 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 it is understood that they do 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 to which 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 the identification and utilisation of lessons by the institution, leading 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 this is entirely feasible. The real challenge is to succeed in convincing actors in the international development community to measure the impact of their programmes 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.

Notes

1 The title of a reggae song by Jimmy Cliff.

2 Of the Office of Evaluation and Oversight of the IADB, I would like to thank for their input particularly Luis Marcano and Pablo Ibarraran 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 came 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 the data were contaminated as they contained identifiable beneficiaries and non-beneficiaries of the programme! Another example is OVE staff were welcomed by an executive agency of a Bank programme; 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. Within OVE the discussion ranged from taking photographs of before and on completion to that only random trials were acceptable. Many of the points discussed paralleled the issues raised in the review of the debate on methodology standards by Coryn (2007).

6 All of OVE's evaluations are made public. In the case of individual programme'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 Standalone programme evaluations were previously studied for two of these themes. See for the housing case Ruprah and Marcano (2007) and, for the 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 standalone evaluation should be undertaken, and then the experience drawn from this should be applied to other similar programmes 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 programme 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 per cent used existing surveys and only 18 per cent used surveys commissioned by OVE.

11 This has required detailed terms of reference that are sent out through EVALNET in which the terms of reference ask for an evaluation with options, i.e. the price of: (i) existing data; or (ii) new data; or (iii) a combination of both. 'Processed' means the evaluation document is reviewed, formatted and put on the web simultaneously with a paper version produced. 'Unprocessed' means such a process has not been completed. All of OVE's evaluations are made public. The Office produces two types of individual project evaluations reports; Working Papers and *Ex Post* Project Evaluation Reports. The latter contains, in addition to the impact results, details of the IADB project (i.e. process evaluation). In the case of individual programme's *ex post* reports and working papers, perhaps due to the Office's under-investment in dissemination, there is an increasing stock of unprocessed reports

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.

References

Benavente, J.M., Crespi, C., Maffioli, A. (2007) *The Impact of National Research Fund: An Evaluation of The Chilean FONDECT*, WP-05/07, OVE, October

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

CGD (2006) 'When Will We Ever Learn? Improving Lives through Impact Evaluation', Washington DC: Center for Global Development

Coryn, Chris L.S. (2007) 'The Holy Trinity of Methodological Rigor: A Sceptical View', *Journal of MutiDisciplinary Evaluation* 4.7

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

Dufflo, E. and Kramer, M. (2005) 'The Use of Randomization in the Evaluation of Development Effectiveness', in G.K. Pitman, O. Feinstein and G. Ingram (eds), *Evaluating Development Effectiveness*, Washington DC: IEG, World Bank

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

Heinrich, C. (2005) *Demand and Supply-Side Determinants of Conditional Cash Transfer Program Effectiveness: Improving the First-Generation Programs*, WP-05-05

Heinrich C. and López, Y. (2005) *Does Community Participation Produce Dividends in Social Investment Fund Projects?*, WP-01-07

IADB (2003) *Ex Post Policy of Operations*, GN-2254-5, September

Ibarraran, P. and Rosas, D. (2006) 'IDB's Job Training Operations: Thematic Report of Impact Evaluations', OVE, IADB, October, 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

Marcano, L. (2005) *Una Evaluación de Impacto del Programa de Fondo de Inversión Social de Panamá*, WP-02-05

Marcano, L. and Ruprah, I.J. (2007) 'Incapacity to Pay or Moral Hazard? Public Mortgage Delinquency Rates In Chile', OVE, October, unprocessed

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

Pritchett, L. (2002) 'It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation', *Public Reform* 5.4: 251–309

Ruprah, I. and Marcano, L. (2007) *A Meta-Impact Evaluation of Social Housing Programs: The Chilean Case*, WP-02-07

Savedoff, W. and Levine, R. (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, H. (2006) I*mpact Evaluation: The Experience of the Independent Evaluation Group of the World Bank*, Washington DC: IEG, World Bank

Wassenich, P., (2006) Peru Rural Road Rehabilitation and Maintenance Project (PE-0136) & National Rural Transportation Infrastructure Program, Stage II (PE-0140)', OVE, May, not processed