A REVIEW OF THE STATE OF IMPACT EVALUATION

Howard White, Shampa Sinha and Ann Flanagan
Independent Evaluation Group, World Bank

1. INTRODUCTION

Impact evaluation is increasingly recognized as an important part of the evaluation toolkit, resulting in calls for 'more and better' impact evaluation. This call has been fuelled by the report of the Centre for Global Development claiming there have been virtually no impact evaluations being undertaken of aid programs at present. But the evaluation departments of the various development agencies have been undertaking evaluations for many years, including studies which cover the question of the impact of the program.

The DAC Evaluation Network has endorsed the call for 'more and better' impact evaluation, hoping to include other agencies in this initiative. The DAC Evaluation Network will join with ECG and UNEG in a 'network of networks' to move the initiative forward. As a first step, agencies were invited to submit recent examples of impact work to IEG for review.

Agencies are engaged in a wide range of evaluations. Many of these are process evaluations, dealing with implementation issues and focusing on inputs, activities and outputs. Several of these studies do assess the issue of impact. But the way in which they do so in a great many of the studies is based on an evaluative assessment of available evidence rather than an approach that could be called rigorous impact evaluation.

However, this is not to say that what we term rigorous impact evaluations are not being undertaken by evaluation departments. Several examples (listed in Annex A) were identified, using a variety of methods. We do not think this is a comprehensive list, and the number of such studies is growing. For example, ADB has formally redefined impact evaluation from being studies with a longer tem time frame to ones which seek to establish a counterfactual. A similar shift has taken place in IEG.

The criterion we used for classifying an impact evaluation as rigorous was that it explicitly considered the issue of the counterfactual and made some attempt to address it. As discussed below, there are additional issues involved, notably overcoming selection bias and contamination problems. A good number of the studies also address these issues.[1]

Hence the claim that there are no impact evaluations being done is an over-statement. But it is also true that there is indeed a challenge of producing more and better impact evaluations. This paper contributes to meeting that challenge by helping set out a possible agenda. First some of the technical issues are reviewed. Then methods for

[1] The World Bank research department is also establishing a database of impact evaluations (http://www1.worldbank.org/prem/poverty/ie/db/evaluationdb.htm). Their criteria for inclusion are stricter, requiring the problem of selection bias to have been addressed. These studies are mostly, but not only, by World Bank staff.

conducting impact evaluations are discussed, including examples from the various development agencies. Finally some ideas on next steps are put forward as a basis for discussion.

2. CONCEPTS IN IMPACT EVALUATION

Before versus after

The simplest approach is to compare the value of the indicator of interest before and after the intervention. However, this only tells us what happened, not why. It is a description of the factual, rather than an analysis of the counterfactual.

The situation before the intervention is not an adequate counterfactual since other things may have changed between before and after which also affect the outcome. Hence the counterfactual is most usually given by a comparison group of non-beneficiaries.

However, sometimes before versus after *is* valid for impact analysis since attribution is very obvious, so that there is no need for a comparison group.[2] An example of this is the Finnish water supply project and its impact on time taken to fetch water. The time households spent in fetching water fell once water pumps were located in the village. There is no other feasible causal factor so the before versus after comparison tells us the actual impact. This point is not always appreciated. A World Bank (not IEG) evaluation used a comparison group to show that rehabilitated schools were in better condition than those which had not been rehabilitated. In that case, a before versus after comparison would have been more meaningful.

Before versus after analysis is even more likely to be applicable at the output level. Many evaluations have to stop at outputs and make an argument based on plausible association for impact. An example is the AusAid project to promote an ambulatory care model for HIV/AIDS in Thailand. Before versus after showed that the model had indeed been adopted.

Providing information up to the output level can sometimes provide a strong basis for inferring outcomes. Two IEG studies – agricultural extension in Kenya and nutrition in Bangladesh – documented implementation problems which would suggest there would be little impact (which was indeed the case). These examples show that a strong factual of what happened can be as an important a part of a good impact evaluation as a solid counterfactual.

[2] The other case in which a comparison group is not required is when a modeling approach is used. Such an approach is more common at the macro level, such as the use of computable general equilibrium (CGE) models to assess the impact of policy change. But modeling may also be used at the project level. An example is FINNIDA's evaluation of a road assistance project in Zambia. A before versus after analysis was used to show how the project had improved road maintenance activities. The actual condition of the road was compared with a forecast of the road's condition under pre-project maintenance levels. The need for a comparison group is apparently side-stepped. However the model needs to be calibrated, and one way of doing this would be using a comparison group. But if that is not available then there may be, as in this case, an existing literature which provides the parameters for the model.

Single and double difference project impact estimates

The difference in outcomes between the treatment group (the beneficiaries) and the comparison group is the most common approach to impact analysis. This is the single difference. Where baseline data are available then the double difference (the difference in the change in the outcome) may be used instead, and in some cases this approach overcomes the problem of selection bias (see below).

An alternative approach is to use a regression in which project participation is an explanatory dummy variable. This approach is equivalent to (i.e. gives identical results to) single or double difference, depending on the specification of the regression. However, other variables may also be included giving a better result, allowing for analysis of differential effects within sub-samples of the treatment group, and removing selection bias under certain conditions.

Although these approaches are very common they are a "black box" approach. They do not identify the channels through which the intervention is having its effect and so are inconsistent with a theory-based approach. These effects are very likely through the determinants which have also been included on the right-hand side – how else will the intervention have an effect other than by affecting the determinants? But if these determinants are included in the regression then the actual impact of the project is under-estimated by the coefficient on the project dummy. But if they are not included then the project dummy may be an over-estimate. Hence it is often better to not use a project dummy but to model the determinants and how the intervention has affected these determinants. This is the 'modeling of determinants' approach described below.

The problem of selection bias

It is usually the case that project beneficiaries have been selected in some way, including self-selection. This selection process means that beneficiaries are <u>not</u> a random sample of the population, so that the comparison group should also not be a random sample of the population as a whole, but rather drawn from a population with the same characteristics as those chosen for the intervention. If project selection is based on observable characteristics then this problem can be handled in a straightforward manner. But it is often argued that unobservables play a role, and if these unobservables are correlated with project outcomes then obtaining unbiased estimates of project impact becomes more problematic.

Two examples illustrate this point:

1. Small businesses which have benefited from a micro-credit scheme are shown to have experienced higher profits than comparable enterprises (similar locations and market access) which did not apply to the scheme. But beneficiaries from the scheme are selected through the screening of applications. Entrepreneurs who make the effort to go through the application process, and whose business plans are sound enough to warrant financing, may anyhow have done better than those who could not be bothered to apply in the first place or whose plans were deemed too weak to be financed.

2. Many community-driven projects such as social funds rely on communities to take the lead in applying for support to undertake community projects, such as rehabilitating the school or building a health clinic. The benefits of such community-driven projects are claimed to include higher social capital. Beneficiary communities are self-selecting, and it would not be at all surprising if those which have higher levels of social capital to start with are more likely to apply. Comparing social capital at the end of the intervention between treatment and comparison communities, and attributing the difference to the intervention, would clearly be mistaken and produce an over-estimate of project impact.

The evaluation design must decide how to handle selection bias. The following steps are a decision tree to assist design (laid out in flow chart form in Annex B):

1. If the evaluation is being design ex-ante, is randomization possible? If the treatment group is chosen at random then a random sample drawn from the sample population is a valid comparison group, and will remain so provided contamination can be avoided. This approach does not mean that targeting is not possible. The random allocation may be to a subgroup of the total population, e.g. from the poorest districts.
2. If not, are all selection determinants observed? If they are, then there are a number of regression-based approaches which can remove the selection bias.
3. If the selection determinants are unobserved then if they are thought to be time invariant then using panel data will remove their influence, so a baseline is essential (or some means of substituting for a baseline).
4. If the study is ex post so a panel is not possible and selection is determined by unobservables, then some means of observing the supposed unobservables should be sought. If that is not the case, then a pipeline approach can be used if there are as yet untreated beneficiaries.
5. If none of the above are possible then the problem of selection bias cannot be addressed. Any impact evaluation will have to rely heavily on the program theory and triangulation to build an argument by plausible association.

The contamination problem

Contamination (or contagion) comes from two possible sources. The first is own-contamination from the intervention itself as a result of spillover effects. To ensure similarity of treatment and comparison groups, a common approach is to draw these groups from the same geographical area as the project. Indeed neighbouring communities, or at least sub-districts, are often used. But the closer the comparison group to the project area the more likely it is to be indirectly affected in some way by the intervention. An agricultural intervention can increase labour demand beyond the confines of the immediate community. There is thus a tension between the desire to be geographically close to ensure similarity of characteristics and the need to be distant enough to avoid spillover effects. Of course, where spillover effects are clearly identifiable they should be included as a project benefit or cost.

But distance will not reduce the possibility of external contamination by other interventions. The desired counterfactual is usually a comparison between the intervention and no intervention. But the selected comparison group may be subject to

similar interventions implemented by different agencies, or even somewhat dissimilar interventions but which affect the same outcomes. Such a comparison group thus gives a counterfactual of a different type of intervention. Different comparison groups may be subject to different interventions. If data are being collected only *ex post,* the presence of similar interventions can be used to rule out an area as being a suitable comparator, though this selection process may leave rather few eligible communities. But in the more desirable situation of collecting baseline data prior to the intervention then there is little the evaluation team can do to prevent other agencies introducing projects into the evaluation comparison area between the time of the baseline and endline surveys.

The first step to tackle the problem of external contamination is to ensure that the survey design collects data on interventions in the comparison group, a detail which is frequently overlooked, thus providing an unknown bias in impact estimates. The second step is to utilize a theory-based approach, rather than a simple with versus without comparison, the former being better able to incorporate different types and levels of intervention.

When there's no baseline

More often than not evaluators are called upon to evaluate a program ex post, and there turns out to be no baseline. Or, if there is a baseline, it was too little (small sample, especially of control if there is one at all) or too late (toward end of the project). The following alternatives may be followed if there is no baseline:

1. If treatment and comparison groups are drawn from the same population and some means is found to address selection bias (which will have to be quasi-experimental, since randomization is ruled out unless the treatment had been randomized, but if the program designers had thought of that they will have thought of a baseline also), then a single difference estimate is in principle valid.

2. Find another data set to serve as a baseline. If there was a baseline survey but with a poor or absent comparison group, then a national survey might be used to create a comparison group using propensity score matching. This method was used by IEG in its analysis of the Bangladesh Integrated Nutrition Project (World Bank, 2005). Or it may be that there was an earlier survey covering both beneficiaries and non-beneficiaries which might be used for evaluation purposes, though it would be very rare to be able to follow up with a second survey and so obtain the panel required for double differencing. Earlier surveys were used to construct the argument in Danida's analysis of the Noakhali Rural Development Project and IEG's analysis of extension services in Kenya.

3. Field a survey using recall on the variables of interest, as was done in IFAD's studies of three West African rural development programs. Many commentators are critical of relying on recall. But all survey questions are recall, so it is a question of degree. The evaluator need use his or her judgment as to what it is reasonable to expect a respondent to remember. It is reasonable to expect people to recall major life changes, introduction of

new farming methods or crops, acquisition of large assets and so on. But not the exact amounts and prices of transactions. When people do recall there may be telescoping (thinking things were more recent than they were), so it is useful to refer to some widely known event as a time benchmark for recall questions.

4. If all the above fail, then the study was make build a strong analysis of the causal chain (program theory). Often a relatively descriptive analysis can identify breaks in the chain and so very plausibly argue there was low impact. In the case of IEG's study of agricultural extension in Kenya it was shown that extension workers spent far less time with farmers than intended and that their extension messages (which did not reflect new research as planned) had often already been adopted. Hence little impact could be expected, and so the evidence of low impact is very plausible even if the comparison group might be faulted on grounds of technical rigor.

5. The argument can be further strengthened by triangulation (indeed this point applies whatever method is adopted): drawing on a variety of data sources and approaches to confirm that a similar result obtains from each. Such an approach is adopted in many of the studies reviewed, most notably the Danish studies of support to rural development in Bangladesh and Mozambique.

3. IMPACT EVALUATION DESIGN

Experimental approaches

Well-designed and well-implemented experimental studies provide a good measure of project impact. By experimentation we mean the random selection of two groups – control and treatment, beneficiaries and non-beneficiaries of an intervention – such that the only difference between the two groups is the variable of interest, i.e. the impact of the intervention. Four examples are given here.

In an evaluation of the impact of corruption on community driven development projects funded by the Kecamantan Development Project (KDP), treatment and control groups were randomly chosen for three different interventions. The first intervention increased the probability of a government audit from four percent to 100 percent. Randomization was by subdistrict so that all villages in a given subdistrict were either audited or not, reducing the possibility of spillover effects. The remaining two interventions were designed to increase community-level participation, increasing accountability and reducing corruption. Contamination was not an issue and villages were randomly selected for treatment. Candidate villages had already submitted projects for approval and been awarded funds for planned road projects however, construction had not yet begun. The treatments were announced to villages prior to the start of construction but after villages had been randomly assigned to one of three treatment groups.

Cambodia's large-scale experiment in contracting with NGOs to provide health care services randomly selected treatment and control villages. Candidate rural villages in Cambodia were surveyed in 1997 and baseline data were obtained on the current state of health care facilities and services, health care coverage and the distributional equity of health care services. Using these data, health care indicators were measured for each district and targets were set for improvement in the indicators. The project mandated that services be targeted to the poor. To avoid the possibility of NGOs selecting out, the project goal and the performance indicators for each district were set in advance and distributed to potential services providers (both NGOs and the district-level government managers) prior to a competitive bidding process. Proposals from NGOs were chosen on quality and cost factors. Ten villages in Cambodia were randomly chosen from the set of candidate rural districts; districts were randomly assigned to treatment and control health care models. The treatment models were contracting-out and contracting-in for service provision.

Attempts were made to ensure that districts were as similar as possible. Districts were of the same size in terms of population and were not receiving additional support for health services from the Ministry of Health as part of another program or government support for a provincial hospital. Districts receiving large amounts of assistance the donor community were also excluded from the sample. To avoid contamination, the districts were spread across three provinces. The baseline data collected was to ensure that the samples chosen were truly random. Candidate rural villages in Cambodia were surveyed in 1997 and baseline data were obtained on the current state of health care facilities and services, health care coverage and the distributional equity of health care services; the same villages were resurveyed in 2001 (two years into a four year program) using the same survey instruments with few exceptions. The data showed the initial distribution of health care services favored the non-poor in nine out of the ten districts.

PROGRESA identified a sample 506 of the poorest communities in Mexico using national census data to construct a wealth index by community and survey data to identify poor households within communities. The sample population also shared other common features such as population size and proximity and access to health and education facilities making the sampling population as similar as possible. Treatment and control groups were then both randomly selected from the sample population. The experimental design of PROGRESA does not allow for quantifiable estimates of behaviors along the causal chain from intervention to outcome. From the sample population of 506 eligible communities, 320 communities were randomly selected and members of the communities received the same package of benefits. Had the experimental design also randomly selected subgroups given partial benefit packages, the partial effects of the education and health aspects of PROGRESA on poverty alleviation could have been estimated (Coady, 2002).

Schools in two cities in India -- Vadodara and Mumbai -- were randomly assigned to treatment and control groups for both remedial education instruction and computer assisted learning interventions. To limit selection bias, the programs were not announced until the beginning of the school year. In education interventions, attrition and lobbying for inclusion in the program can introduce contamination and bias into the evaluation. A PAL evaluation of the experiment suggested that parents are unlikely to seek information on programs available in schools reducing the possibility

of lobbying for inclusion. Further, administrators are unlikely to grant transfers and there are a limited number of options available for students to learn in a given area and a given language greatly reducing bias from attrition (Banerjee, Cole, Duflo and Linden, 2006). Attrition rates were checked and were similar across treatment and control groups and there was no evidence that students leaving either program had similar characteristics, i.e. were self-selecting.

Natural experiments

Natural experiments occur when participant is allocated in a way which is not at all correlated with expected outcomes. In that case a sample of non-beneficiaries will be a valid control group. One example of a natural experiment is class size in Israel (Angrist and Lavy, 1999). There is a debate on the impact of class size on student learning. But class size may be endogenous with respect to other factors influencing outcomes, such as school management. However, in Israel class size is exogenous since by law no class can exceed 40. Hence once there is a 41st pupil the class is split into two class of 20 and 21 each. Another example comes from land titling in Argentina. Squatters outside Buenos Aires were awarded title to the land on which they were squatting with compensation paid to the original owners. Some owners disputed the settlement in court, so these squatters did not obtain the land title. Which squatters got title or not had nothing to do with the characteristics of the squatters. Hence non-title holders and title holders can be compared to examine the impact of having title on access to credit (there was none) and investing in the home (there was some).

Pipeline approach

In the pipeline approach communities, households or firms selected for project participation, but not yet treated, are chosen for the control. Since they have also been selected for treatment there should in principle be no selectivity bias, though there may be. For example, if the project is treating the "most eligible" first then these units will indeed be systematically different from those treated later. If this is the case, then the approach ensures a bias rather than avoids it. For example, phase one may start with the poorest families or alternatively with the more centrally located or better-off areas, and in both of these cases the characteristics of communities in later phases are likely to be different.[3] This approach also assumes that there has been no change in selection criteria. This is why project design and selection criteria must be carefully reviewed when applying this approach because there will often be systematic differences between the phases. Clearly, the approach can only be used for activities which continue beyond the end of the project being evaluated. The data on the pipeline group can also serve as a baseline in future studies and therefore help to establish an efficient impact evaluation system.

Two examples of the pipeline approach come from the evaluation of microfinance programmes in Nigeria, Malawi, Haiti and Kenya conducted by UNCDF and in Pakistan by DFID. The UNCDF evaluation followed the pipeline approach more

[3] IEG tried to use a pipeline comparison group to evaluate an irrigation project in Andhra Pradesh, India but found that farmers covered by the later phases were typically more remote and different in other ways from phase one farmers (White, 2006).

closely by using new clients as the control group and mature and ex clients as the treatment group. New clients were defined as those who had not yet received their first loan or who had received their first loan but had yet to finish their loan cycle. The treatment group comprised of two-year clients or clients who had joined the programme at least 20 months prior to the survey. The control and treatment groups were further disaggregated by location (rural, urban, peri-urban) and sector of economic activity (trade, agro-business, service and manufacturing) and impact assessed at the individual, household and institutional level using a range of qualitative (e.g. empowerment, client self-esteem) and quantitative (e.g. household asset acquisition) variables. The results obtained from disaggregation enabled the evaluators to make recommendations in favor of market segmentation.

In the Pakistan case, the evaluators were prevented from using new clients who had not been "treated" at all as a control group since loan disbursement by the implementing NGO (Kashf) is so quick—sometimes within 24 hours—that there was not a large enough group of people still waiting for their first loan. Consequently the control group was composed of people who had been with the programme for less than six months and were in their first loan cycle.[4] In designing the sampling frame steps were also taken to ensure that early and late entrants to the programme had had an equal amount of exposure to eliminate the need to investigate why people joined the programme at different stages, which would in turn entail the need to control for differences in the propensity to take risks. Therefore, new clients were selected from new branches, while mature clients were chosen from branches which were three years or older.

A final example of a planned pipeline approach is IDB's evaluation of vocational training in Panama, in which future selected participants were used as the comparison group. However, because of financial constraints this second group never received the training, so that the design might be called a natural experiment.

Quasi-experimental approaches

Propensity score matching

Selection may be based on a set of characteristics rather than just one. Hence the comparison group needs to be matched on all these characteristics. This may seem a rather difficult task. But it can be managed through a technique called propensity score matching (PSM). Once the control is identified then project impact can be estimated using single or double difference estimates.

Propensity score matching can be attractive for two reasons. First, comparison group data may have been collected but are thought not to be representative because of selection bias. Second, there may be data only on the treatment group but not the control. A different, possibly nationwide, data set can then be used to construct a comparison group using PSM.

[4] Initially the evaluators tried using non-participants living in the same area as the participants as a control group but abandoned this method since it did not allow them to compare "like with like" and, in the absence of panel data, it was unable to fully capture the changes over time that could be attributed to the programme.

The steps involved in carrying out propensity score matching are as follows:

1. Obtain a control dataset.
2. Run a participation model (probit/logit regression).
3. Calculate participation probabilities.
4. Drop observations outside the region of common support (i.e. observations in the treatment group whose probability of participation exceeds that of any from the potential comparison group, or those from the latter group with participation probabilities below those of any members of the treatment group).
5. Match observations based on participation probabilities.
6. Calculate project effect for each pair (or set) of matched observations.
7. Calculate the average of these differences (project effect).

Examples of propensity score matching are:

1. IDB's study of vocational training in Chile used propensity score matching to obtain a comparison group, since the universe of non-participants may well have different characteristics, especially given the program's screening process.

2. This approach was also used in the four IDB studies of support to science and technology, specifically grants to academics. The problem of selection bias is very clear in this case. The awards were given on a competitive basis, so that the performance (measured by publications) of those selected should have been better than those who were not selected even in the absence of the program. Hence a propensity score was calculated based on variables likely to affect success, such as the quality of the applicant's own educational institution and their prior publication record. A positive impact was found.

3. The District Primary Education Program (DPEP) in India put a lot of emphasis on M&E. But it only collected data on project districts, so no impact evaluation was possible. Census or household survey (NSSO) data allow for some comparisons, though on a more limited range of outcomes than desirable, and only allowing counterfactual analysis at the upper (outcome) end of the log frame. But DPEP districts were chosen based on low female literacy and the potential to improve. Hence DPEP districts cannot be compared with non-DPEP districts in a straightforward manner. An impact analysis by the World Bank used propensity score matching to match districts (always matching a district with one in the same state), modelling program participation on female literacy (overall and among Scheduled Castes and Tribes), proportion of STCs, population density, housing quality and village infrastructure. The results show an improvement in enrolments and progression beyond primary school, especially for minority groups.

4. The Bangladesh Integrated Nutrition Project commissioned evaluation surveys from six project sub-districts and two control areas. The sample size for the controls was rather small, and since they were contiguous with the project districts there was a likelihood that spillover effects (a major focus of the project was nutritional counselling—word can spread) would reduce measured

impact in the project versus control comparison. A World Bank study used a national nutrition survey to create a control group using propensity score matching. A comparison of the findings shows that this approach yields more internally consistent results—finding the impact to be very low for the money spent. A food subsidy to all households with children would have yielded a larger nutritional impact than did the project-supported activities.

Regression discontinuity

Regression discontinuity uses the propensity score in another way. The outcome variable is regressed upon the score including a program dummy (possibly both intercept and slope, as in equation 5). The fitted values are calculated using the mean score for the treated and both D=0 and D=1. The difference between these two fitted values is the program impact. This method was used in IDB's study of support for scientific research in Chile (this study also used propensity score matching and was listed above).

Modeling the theory

The above approaches give an estimate of impact, but may give no indication as to the channels through which this impact has been felt. The alternative approach, currently being used by IEG in its impact evaluations, is to model the determinants of the outcomes using regression models. The determinants of these determinants are also modeled, working down the results chain until the link is made to program inputs.

4. NEXT STEPS

The proposed next steps for this group are:

1. Agree the basics for the proposed "Guidelines for Impact Evaluation" and designate working group to prepare draft to be submitted to the next meeting.
2. Draw up a program of impact evaluation work, including the possibility of joint impact evaluation. This program should identify areas in which there are few studies.
3. Explore activities to build impact evaluation capacity within agencies and within developing countries.
4. Examine how to extend impact analysis into other products, such as country evaluations, and for other aid instruments, such as sector and budget support.

References

Angrist, Joshua D., and Victor Lavy, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," Quarterly Journal of Economics 114 (May 1999), 533-575.

Banerjee, Abhijit, Shawn Cole, Esther Duflo and Leigh Linden. 92006). "Remedying Education: Evidence From Two Randomized Experiments in India," Center for Economic Policy and Research, Discussion Paper Series No. 5446.

Coady, David (2003). "Alleviating Structural Poverty in Developing Countries: The Approach of PROGRESA in Mexico," International Food Policy Research Institute, Washington, DC

Galiani, Sebastian and Ernesto Schargrodsky, “Property Rights for the Poor: Effects of Land Titling” draft, March 13, 2006 (Available at www.tinyurl.com/ndw69).

Olken, Benjamin (2005). "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," National Bureau of Economic Research.

Rawlings, Laura, Lynne Darling Sherburne-Benz and Julie VanDomelen, " Evaluating Social Funds: A Cross-Country Analysis of Community Investments (Regional and Sectoral Studies)" (World Bank: 2004)

Schwartz, J. Brad and Indu Bhushan. "Cambodia: Using Contracting to Reduce Inequity in Primary Health Care Delivery,"

Annex A List of impact evaluations

Agency	Date	Project	Country	Project description	Project Objective	Major findings	Data	Methodology
AfDB	2005	Electricity sector	Egypt	13 projects and 2 studies supporting electrification in the country, which accounted for 31% installed generating capacity and connected 200 villages.	To support government's efforts to meet largely peak demand through increasing the electricity generation capacity, expanding and strengthening the transmission lines for the unified system and providing rural electrification to a number of villages.	Contributed to meeting increasing demand, with 100% urban coverage and 95% in rural areas. Undoubtedly contributed to socio-economic development, the economic benefits being valued at 4% of GDP.	There are no survey data linked to any of the projects.	The value of electricity output is taken from an Energy Sector Study for Egypt, Morocco and Tunisia.
AusAid	2005	HIV/AIDS Ambulatory Care Project	Thailand	Establishment of a fully integrated ambulatory care model incorporating human resources development, health education and information services, development of organizational capacity and effective project management at the Bamrasnaradura Institute.	To assist Thailand's premiere infectious diseases hospital to develop its HIV/AIDS services and to provide education and training for clinical staff at the hospital and beyond.	1. The ambulatory care model developed continues to operate successfully and to provide significant benefits. 2. There is clear evidence that the training has improved participants' knowledge, attitudes and professional practices.	Document review, observation, statistical analysis and individual and focus group interviews. A survey was conducted of 742 sites throughout Thailand which had received a training package	Chronological pre-post data from the same institution was used as a point of comparison. Purposive sampling techniques were used to assess sustainability.

Agency	Date	Project	Country	Project description	Project Objective	Major findings	Data	Methodology
CIDA	1998	Ghana Water Program	Ghana	Evaluation of CIDA's water sector portfolio in Ghana between 1973 and 1997 with commitments totalling $134 million. The projects were varied, including 1. engineering 2. borehole drilling 3. handpump installation 4. health and hygiene education 5. community animation 6. local maintenance training. 7. community management structures.	To improve health and productivity in Upper and Northern Ghana through improved water supply systems.	Outputs achieved included: 1. Improved knowledge and skills of over 15,000 women in water and health. 2. Increased knowledge of over 3,000 men in pumpsite management and handpump maintenance. 3. Community-based involvement through Pump Management Committees(2,500), Water and Sanitation Development Boards. The goal-level results were listed as 1. eradication of guinea worm 2. increased health awareness as measured by villager testimony and willingness to pay for clean water.	Document review and key person interviews. Internal benchmarking data from project reviews used for before versus after. Baseline data on household health status and water use attituded from three separate water supply projects were available but limited resources did not allow ex post surveys.	In most cases, before versus after with distinct project level goals. For broader developmental impacts, for example poverty reduction, no controls were used to isolate the impact of CIDA's water supply program from other external influences.
DANIDA	2001	Noakhali Rural Development Project	Bangladesh	Two-phase project: Phase 1 aimed at promoting economic growth and social progress, particularly for the poorer sections of the population. NRDP-1 consisted of activities in four areas namely, infrastructure (roads, canals, market places, public facilities), agriculture (credit, cooperatives, irrigation, extension, marketing), other productive activities (livestock, fish ponds, cottage industries), social sector (health & family planning, education.). Particular activities targeting the poorer sections of the	The main objective of the first two phases was emergency agricultural input supply (mainly seed, fertilizer and hand tools) to small farmers. Subsequently the emphasis was on building capacity and establishment of an extension service, thereby improving the living standards of the rural population in Tete Province.	Overall conclusion: "On the whole it must be concluded that although NRDP has played an important role, the intervention has not been on such a scale that it has changed the socio-economic development of entire villages."	A survey was undertaken of the infrastructure. Interviews were conducted with beneficiaries and non-beneficiaries, the latter in both project and non-project villages. Several villages had been previously surveyed in various studies.	Ex-post mixed-methods : quantitative and qualitative with a substantial fieldwork component lasting 4 months (23 personnel). Quantitative methods included quantitative analysis of project monitoring and contextual data. Qualitative methods included documentary study, archival work, questionnaire surveys, stakeholder and informant interviews, representative surveys of project components, assessment of buildings, roads and irrigation canals (function, maintenance, etc), village surveys and interviews, observation, focus groups, case-studies.

Agency	Date	Project	Country	Project description	Project Objective	Major findings	Data	Methodology
				population consisted of the creation of temporary employment in construction activities and engaging them in income generating activities.				
DANIDA	2002	Agricultural Development Project	Mozambique	The project ran from 1985-1999 and went through multiple phases – starting as an emergency supply project (mainly in the town of Tete), slowly developing into an agricultural support project in three distinct districts and ending as an institutional support project enabling the Ministry of Agriculture to service farmers.	The main objective of the first two phases was emergency agricultural input supply (mainly seed, fertilizer and hand tools) to small farmers. Subsequently the emphasis was on building capacity and establishment of an extension service, thereby improving the living standards of the rural population in Tete Province.	1. The project provided modest support for poverty reduction 2. Social processes and political context affected the way the project was executed. 3. The project may have made some contribution to an alteration in the tense relationship between the Government and farmers	Data were collected via documentation reviews (including mid-term review documents), other socio-economic studies, surveys, interviews.	There were unique methodological difficulties in establishing a pre-post counterfactual due to the civil war ending in 1992. There was no baseline study available of the pre-project situation. Since living standards rose after the end of the war, this also made attribution problematic.

Agency	Date	Project	Country	Project description	Project Objective	Major findings	Data	Methodology
DFID	2004		Pakistan	NGO microfinance and enterprise development programme.		Kashf was found to be efficient in targeting the poor: 90 percent of its clients were living on less than $1 a day. The income of Kashf's mature clients is 51 percent higher than its new clients, suggesting that Kashf is helping alleviate poverty. Furthermore, the longer a client is associated with Kashf the higher their income is.	Survey using new participants as comparison group.	A pipeline approach was planned. But Kashf loan disbursement is so quick that there was not a large enough group of people still waiting for their first loan. Therefore, the control group was composed of people who had been with the programme for less than six months and were in their first loan cycle.
DFID and World Bank (study financed under DFID-World Bank Strategic Poverty Partnership Trust Fund)	2005	Kecamatan Development Project		CDD project for village infrastructure. This study was concerned with limiting corruption in the use of funds.		Audits reduced the level of unaccounted funds by 8 percent, more than enough to offset the cost of the audit. By comparison, increasing grass-roots participation in the monitoring process only reduced the discrepancy between funds budgeted for wages and those actually received; there was no impact on the theft of funds destined to be spent on road-building materials. Since materials account for three-quarters of road-building expenditures, increasing grass-roots participation had little impact overall.	Comparison of estimates of the actual cost of the materials used to build roads—based on analysis of core samples from each road—with the village account books, which itemize all road-building expenditures. Stolen funds must therefore show up in the difference between reported expenditures and estimated actual expenditures.	A randomized field experiment was conducted in 608 Indonesian villages. A random sample of villages was chosen to be audited by a central government agency. Government audits rarely result in prosecution but because the results are divulged in open village meetings they can lead to substantial social sanctions. An analysis was also made of grass-roots monitoring. In the first experiment villagers were encouraged to take part in village accountability meetings. In the second experiment, villagers were given the opportunity to fill out a form assessing a village project. The responses were summarized and read out at a village meeting.

Agency	Date	Project	Country	Project description	Project Objective	Major findings	Data	Methodology
EBRD	2006	Support to MSMEs	Bulgaria, Georgia, Russia and Ukraine	Loans tailored to needs of borrowing firm, supported by TA	Provide loans to firms which may not otherwise have access to finance.	Positive effect of loans on a range of outcome variables.	1,272 MSMEs (defined as firms with fewer than 250 employees) in Bulgaria, Georgia, Ukraine, and Russia. Treated (received loan in 2002) and control, latter having received non-EBRD loan in 2002.	Careful matching on pre-2002 characteristics; panel data analysis. Also make IV estimates to allow for endogeneity of receiving a loan.
EC	2000	North-South Cooperation Against the Spread of HIV/AIDS and Assistance for Demographic Policies and Programs						Analysis of existing documents, interviews, field visits in 10 countries.
FINNIDA	1996	Road Assistance Project: Phase II	Zambia	Mainly technical assistance for the development of improved systems by means of practical and effective training, with some funds for parts and equipment	Strengthen sustainable maintenance capabilities of the Roads Department, with an emphasis on development of improved systems by means of practical and effective training	Project prevented 36 km of road from falling into total disrepair, but was unable to prevent deterioration of roads of good quality before the project. Equipment availability and staff morale found to have improved.	Surveys of road quality undertaken in 1992 and 1995.	Expected road quality in 1995 was projected from 1992 data assuming without project levels of maintenance, and these compared with the 1995 outturns.
FINNIDA	2001	Finland's Support to Water Supply and Sanitation	Multi-country	Installing hand pumps and related institutional development activities	Not explicitly stated: appears to be providing water supply	One of the greatest benefits has been improved time saving.	Surveys undertaken in country studies	Before versus after: no counterfactual us need

Agency	Date	Project	Country	Project description	Project Objective	Major findings	Data	Methodology
FORMIN	2006	Impact Assessment Report of the Soil and Water Conservation Projects (SWCP) of EECMY-LWF/WS in Southern Ethiopia	Ethiopia	Assessment of four sites where Soil and Water Conservation Projects were implemented: Otore Harre river diversion, Lega Surre Wurwita earthen micro dam, Furuna river diversion and Lower Bilate river diversion.				Participatory rural appraisal.
IDB	2006	IDB's Science and Technology Programs: An Evlauation of the Technology Development Funds and Competitive Research Grants	Chile, Columbia, Uruguay, Brazil, Argentina and Panama	A multi-country evaluation of the impacts of two science and technology investment instruments: Technology Development Funds (to spur productive innovation) and Competitive Research Grants (funding basic R&D).	Regain competitiveness in the new global and open economic environment.	Findings on 1. crowding out 2. innovative outputs 3. firm performance 4. scientific production varied by project/country and controls used in regressions. At the program level the evaluation found no evidence of crowding out. The results for innovations were not significant (no effect). Mixed results (by project/coutnry) for productivity, employment and sales; number of publications and citations; change in publications and citations.	1. Project-level data (grouped by country)	A variety of methods were used in the evaluation. 1. Single difference with propensity score matching 2. Doube difference with and without propensity score matching 3. Panel data fixed effect IV estimation (beneficiaries and non-beneficiaries), and 4. discontinuity regression.
IDB	2006	FONTEC Program (financing innovation in the private sector) and FONDECYT (financing basic and applied research	Chile	R&D support through public technology funds.	To increase the competitiveness of Chilean economy through technological innovation in strategic productive sectors, in particular SMEs.	Positive and statistically significant effect of FONTEC on the level of investment in R&D (pesos) but no effect on R&D intensity (R&D as a percent of sales). No statistically significant effects were found relating FONDECYT to increases in either the number or	Survey data on 219 beneficiary firms and 220 non-beneficiary firms representing the sectoral and geographical distribution of FONECT recipient firms. FONDECYT	Double-difference with propensity score matching. Regression discontinuity.

Agency	Date	Project	Country	Project description	Project Objective	Major findings	Data	Methodology
		activities).				quality of publications.	specific database on all projects financed between 1998 and 2004 including bibliometric information attributable to the program. A control group was chosen from among the group of rejected proposals.	
IDB	2006	FONTAR, Technical Upgrading Program	Argentina	Financial support for scientific and technological R&D.	To increase efficiency, productivity and competitiveness through technological upgrading and support for innovation processes, especially for SMEs.	FONTAR program increased R&D intensity and private R&D expenditures of new firms. No significant effect found for innovative output, labor productivity or export capacity.	Survey data for 414 firms (1998, 2001-2004) of which 136 firms applied for and received FONTAR subsidies. 62 firms not granted subsidies and 216 non-participating firms (no application made).	Double-difference with propensity score matching.
IDB	2006	PNDCyT financing for basic and applied research.	Columbia	Financial support for science and technology sector.	To strengthen Colombia's capacity to conduct scientific research through infrastructure and training to facilitate sustainable development.	Significant increases in the number and quality of publications for financed researchers. Financed researchers published 1.18 times the number of publication of non-financed researchers. Higher quality of research, researcher age and institutional rank were also positively related to number of publications.	Administrative data on PNDCyT program and secondary bibliometric data sources. Registry of researchers used to form comparison group.	Regression with controls, single-difference and propensity score matching.

Agency	Date	Project	Country	Project description	Project Objective	Major findings	Data	Methodology
IDB	2005	PROJoven (Youth vocational training)	Peru		The overall objective of the program was to help economically disadvantaged youths, of 16-24 years old, to enter into the formal labor market by providing them with training and an opportunity to acquire work experience, which is based on the needs of the private sector.	There are positive and statistically significant effects in terms of paid jobs and formal employment probabilities, as well as in terms of monthly earnings. We also find that female youngsters and 16-20 year olds seem to benefit more from the program. In general, they experienced higher PROJoven impacts on paid job probabilities, formal jobs probabilities and monthly earnings than their male and 21-25 year olds counterparts.	Project dataset covering both beneficiaries and control. The selection of the control group is based on the following variables: age, sex, education, poverty level and geographic residence.	Propensity score matching using a difference model, so also controls for time invariant unobservables.
IDB	2006	PROCAJOVEN (job training program).	Panama		The program's objective was to improve labor market prospects for unemployed and disadvantaged youth aged 18-29 by providing job readiness and technical training.	Significant increases in both employment rates and earnings for women and participants in the Panama region of 12 to 15 percentage points.	A subsample of the 3,700 PROCAJOVEN beneficiaries and a subsample eligible participants not receiving training (due to funding constraints) stratified by age, gender and educational class.	Pipeline which turned to natural experiment.
IFAD		Rural Finance and Community Initiatives Project	The Gambia				(i) Quantitative survey of 253 households and 72 kafos; (ii) a qualitative PRA exercise conducted for eight kafos.	Double-difference: as no baseline survey had been conducted, the survey had to adopt the 'recall method' in order to obtain an estimate of corresponding values in 2000 for income, assets, savings and employment.

Agency	Date	Project	Country	Project description	Project Objective	Major findings	Data	Methodology
IFAD		Upper East Region Land Conservation and Smallholder Rehabilitation Project (LACOSREP) – Phase II	Ghana			Households served by the project had significantly increased their assets. Those households that received financial services from participating banks have reported benefits in terms of better opportunities for investments in trading and farming, but also in activities that do not directly generate income but contribute to household welfare such as health, schooling and the situation of women. All the beneficiaries interviewed acknowledged that the project has enhanced their food security. Women were not traditionally land owners in this region, but the WUA system has given them direct access to irrigated land. The overall impact of LACOSREP II on beneficiary communities has been considerable in the areas of food security, income generation, cohesion, literacy and promotion of gender issues. The achievements of LACOSREP should be seen against the trend of increasing poverty and environmental degradation in the UER.	(ii) a preliminary qualitative field survey of Water User Associations (WUAs); (iii) a preliminary quantitative field survey of 189 households (159 beneficiary and 30 non-beneficiary) comparing a before and after situation (using recall method); (iv) an ad hoc quantitative survey of non-beneficiaries; and (v) participant observations and interviews as recorded by mission members	Both qualitative and quantitative techniques have been adopted and the main findings presented in this report are results of triangulation between different methods and sources. Due to significant delays in the implementation of the water management component, the analysis of impact and sustainability is largely based on longitudinal data and observations of LACOSREP I sites as well as inferences from the early data and observations from the few completed LACOSREP II dams.

Agency	Date	Project	Country	Project description	Project Objective	Major findings	Data	Methodology
IFAD		Upper West Agricultural Development Project	Ghana			The overall impact of UWADEP has been quite modest, due to limited implementation achievements. Few irrigation infrastructure facilities were completed and functional by project closure, making it difficult to assess impact properly. Those households that have received financial services from participating banks have reported income increases through opportunities for investments in trading and farming.	Findings are the result of triangulation between multiple sources including: (i) quantitative surveys of beneficiaries and non-beneficiaries; (ii) a qualitative survey of five dam sites.	Double difference using recall.
JBIC	2006	Jamuna Multipurpose Bridge Project (JMBP)	Bangladesh	A bridge was constructed across the Jamuna River providing a rail and road link between the less-developed Northwest region and the more developed eastern part of the country that includes Dhaka.	To provide opportunities for economic growth and poverty reduction.	Fishpond: 1. Difficult to attribute the boost in fishpond production entirely to NRDP since market forces played a greater role.	Panel data from 62 villages covering Divided into treatment and control. Baseline data was collected from these villages before the bridge started operating. Open-ended questionnaires in focus groups were also used to provide qualitative data.	Difference-in-difference analysis was done. There was some attrition in sample size over time.
JICA	2006	The Improvement of Teaching Methods in Mathematics (PROMETAM)	Honduras	Development of teaching materials and in-service teaching training.		1. Teachers' test scores improve by 24 points (on scale of 100) from 2002 to 2005; 2. Pupil scores ambiguous impact	Survey of 128 teachers and 404 fourth grade pupils	Single difference before versus after

Agency	Date	Project	Country	Project description	Project Objective	Major findings	Data	Methodology
JICA	1996	Ubon Institute for Skill Development (UBISD)	Thailand	Support for vocational training		Positive rate of return based on benefits calculated as wage differential between UBISD graduates and the minimum wage. The positive impact was weaker for graduates working in Bangkok compared to those in the vicinity of Ubon.	Survey of UBISD graduates. Random sample of individuals = 8 survey responses of 176 surveys sent; "selected companies who hire UBISD graduates" which results in skewed sample (given the need to secure a large enough sample size) = 91 survey responses of an unknown number sent.	Wage differential is between graduate wage and the minimum wage (i.e. single difference with a control taken from national sources). Appropriate adjustments are made for working time and unemployment. This calculation gives the private rate of return. The social return is harder to determine as it requires the productivity differential between UBISD-trained employees and those who would have been employed in their absence.
JICA	2003	SMASSE Project: Instruments for Internal Monitoring and Evaluation	Kenya	A description of the monitoring and evaluation instruments used in the Strengthening of Mathematics and Science Secondary Education (SMASSEE) project.	To improve the science and mathematics abilities of Kenyan students through in-service teacher training.			1. Pre- and post-testing of teachers participating in INSET training 2. Classroom observation 3. Lesson observation 4. Teacher mastery test 5. Measurment of teaching techniques ASEI (Activity-focused, student-centered learning, experiments and improvisation) and Plan, Do, See Improve (PSDI) 6. Achievement level test.
UNCDF	2003	Microfinance Programme Impact Assessment	Nigeria, Malawi, Haiti and Kenya	UNCDF made a policy shift in 1999 that directed the microfinance programme to help microfinance institutions to achieve institutional sustainability, with a focus on financial sustainability	To increase outreach and clients' asset base and enhance the consumption-smoothing potential of poor people.		Individual and focus group interviews, exit surveys, loans and savings information.	

Annex B Decision Tree for Selecting Evaluation Design to Deal with Selection Bias