IEG
Independent Evaluation Group

IMPACT EVALUATION- THE EXPERIENCE OF THE INDEPENDENT EVALUATION GROUP OF THE WORLD BANK

IMPACT EVALUATION- THE EXPERIENCE OF THE INDEPENDENT EVALUATION GROUP OF THE WORLD BANK

Independent Evaluation Group,
The World Bank

Acknowledgement

This report has been prepared by Howard White (IEGSG) under the supervision of Alain Barbu. Comments were provided by Nina Blöndal, Jorge García-García, Gaamaa Hishigsuren, Miguel Laric, Edoardo Masset, Keith Mackay and Patrick Grasso. This booklet was partly produced with support from the IEG-DFID partnership agreement.

Alain Barbu
Manager
Sector, Thematic & Global Evaluation
Independent Evaluation Group

Table of Contents

1. What is Impact Evaluation?

Introduction

Impact evaluation is an assessment of the impact of an intervention on final welfare outcomes.[1] The results agenda has forced agencies to demonstrate that the money they spend is improving the lives of poor people, thereby increasing demand for impact evaluation. In the current environment, calls for increased aid spending are only credible if it can be shown that current spending is indeed contributing toward the attainment of the Millennium Development Goals.

However, the meaning of impact evaluation has taken on different meanings over time, and there continue debates as to how it should be done. This introductory chapter has the following purposes. First, it puts forward the definition of impact evaluation as a 'counterfactual analysis of the impact of an intervention on final welfare outcomes.' Second, it discusses two sources of bias which can result in impact evaluation studies giving misleading results: (1) contagion, and (2) self-selection bias.

Different meanings of impact evaluation

Impact evaluation has taken different meanings during the last twenty years. The following have been the most common:

- An evaluation which looks at the impact of an intervention on final welfare outcomes, rather than only at project outputs, or a *process evaluation* which focuses on implementation;
- An evaluation concerned with establishing the counterfactual, i.e. the difference the project made (how indicators behaved with the project compared to how they would have been without it);
- An evaluation carried out some time (five to ten years) after the intervention has been completed so as to allow time for impact to appear; and
- An evaluation considering all interventions within a given sector or geographical area.

These four definitions are not mutually exclusive – many IEG impact evaluations in the 1990s were carried out five or more years after the intervention closed and also tried to establish a counterfactual. But nor need they coincide. Participatory impact evaluation follows the first definition, reporting beneficiary perspectives on how the intervention changed their lives, but makes no formal reference to a counterfactual.

[1] Intervention can refer to a project, program or policy. However impact evaluation of national-level policy changes requires a different approach to those discussed in this brochure.

IEG's current approach combines the first two definitions, that is a counterfactual analysis and a concern with final welfare outcomes. The rest of this chapter discusses the challenge of establishing the counterfactual.

Debates in impact evaluation

Impact evaluation is the counterfactual analysis of the impact of an intervention on final welfare outcomes. Carrying out such analysis requires analysis of data collected at the appropriate time and places, often using statistical methods which may strike many as unnecessarily sophisticated. Moreover, this technical sophistication may seem to come at the expense of relevance. But there is not necessarily a trade off between what is called here rigorous impact evaluation – i.e. one which applies the appropriate technical procedures - and relevance, which depends on a well-contextualized theory-based approach. These debates about impact evaluation are themselves a context which need be understood to appreciate the different approaches to measuring impact.

Debates over impact evaluation reflect the more general debate over the relative roles of qualitative and quantitative methods in social research. Participatory impact evaluation grew rapidly in the 1980s, and is still going strong especially amongst non-governmental organizations (NGOs). The proponents of the participatory approach are skeptical of the econometricians' attempts to reduce the impact of complex social interventions to a single number. But the econometricians reject analyses which fail to build on a well-designed sample of project and comparison groups which allow statements to be made with a degree of scientific confidence about the behavior of indicators with versus without the intervention. The reports and literature from these different approaches are in general developing in parallel, with rare attempts at dialogue to establish common ground let alone methodological fusion. IEG's position, argued for here, is that a mixed-methods approach produces the strongest evaluative findings, combining well-contextualized studies with quantitative rigor.

A second area of debate is that over the applicability of experimental approaches. This approach is discussed more fully in chapter 2. In brief, the experimental approach requires the random allocation of project participation. Critics argue that the approach is simply not practicable for a broad range of development interventions. However, these critics are stuck with the problem of showing that their alternative, quasi-experimental, approaches solve the evaluation problem (establishing a valid counterfactual) in the way that randomization clearly does. IEG has of necessity been a user of quasi-experimental approaches. The case studies presented in the second half of this volume demonstrate how good contextualization can build confidence around the findings from these approaches.

The impact evaluation challenge: the search for the counterfactual

The counterfactual is a comparison between what actually happened and what would have happened in the absence of the intervention. Data can be collected on the factual. But we cannot observe what would have happened to those affected by the intervention if the intervention had not happened.[2] One solution to this problem has been the before versus after approach. The mean outcomes for the treatment group are compared before and after the intervention and any change attributed to the intervention. Whilst data on treatment group trends are useful information to have, they tell us what has happened to the treatment group and nothing more. It is not possible to attribute the observed changes to the intervention since other external factors may have been partly or wholly responsible for the change, or may even act to offset the positive impacts of the intervention. For example, under-five mortality has been rising in several African countries on account of HIV/AIDS despite increased immunization coverage and access to safe water. So before versus after approach yields either an overestimate or an underestimate, but we will not know which it is.

The better solution to the problem of the counterfactual is to select a comparison group, i.e. a set of individuals, households, firms or whatever, who are like the treatment group in every way, except that they were not subject to the intervention. This sounds like a difficult, though not impossible, task. There are, however, two problems: contamination and sample selection bias.[3]

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 neighboring 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 labor demand beyond the confines of the immediate community. In IEG's study of an irrigation project in Pakistan (see appendix II, case study 1) the comparison area was moved as it was thought to be too close to the intervention area and so subject to demonstration effects.[4] As another example, information on health and nutrition can find its way to neighboring communities by word of mouth. For example, several studies of the Bangladesh Integrated Nutrition Project (discussed in Chapter 6) have found it to have had little or no impact. Defenders of the project attribute this finding to contamination since the control groups from many of the studies were adjacent to the project areas (Levinson and Rohde, 2005; and Sack et

[2] A formal mathematical statement of the problem is given in Appendix I.
[3] Other issues in selecting the comparison group are discussed in Bamberger (2006).
[4] However the evaluation found no difference between project and control after the intervention so that contamination may have occurred nonetheless; or of course the intervention had no impact.

al., 2005). 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.[5]

But distance will not reduce the possibility of external contamination by other interventions. The desired counterfactual is usually a comparison between the zintervention 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 comparison, though this selection process may leave rather few communities as being eligible. But where baseline data have been collected, there is probably little the evaluation team can do to prevent other agencies introducing projects into the 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, which is better able to incorporate different types and levels of intervention.

Sample 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 not 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.[6] 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 microcredit 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

[5] Of course where spillover effects are clearly identifiable they should be included as a project benefit. IEG's study of urban development in Kenya (Appendix II, case study 7) noted spillover benefits in the non-project control from new health and education facilities in the project area.
[6] The problem then becomes a straightforward one of endogeneity to be handled by instrumental variable estimation. In cases where a regression need be estimated just for the treated, then a two part sample selection model (Heckman procedure) is appropriate.

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.

These problems make the attraction of random allocation clear. 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.[7] An alternative is a 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 neither randomization, nor a pipeline approach, are possible then modeling the selection process can help control for selection bias. If selection is entirely on observables, e.g. illiterate women, then selection bias can be completely eliminated. But if there is an important unobservable component to the selection process then, especially if only endline data are available, the danger of bias remains strong. Resorting to a theory-based approach to tell a convincing story can, however, help allay some fears.

If the direction of bias is known then impact estimates can be reported as either upper or lower limits, as was done in an IEG study of rural development projects in Sri Lanka (Appendix II, case study 3). Before versus after yields in the project area were used for one estimate of the rate of the return. Since yields would have risen in the absence of the project (as they did in the non-project areas) this rate of return was an over-estimate of the actual impact, and so provides an upper limit. The second estimate was calculated as a comparison between project and a comparison group of non-project areas. The comparison group was known to be contaminated by interventions supported by other agencies, so the estimate calculated in this way is an under-estimate of project impact, giving a lower limit. Hence the actual rate of return lay between these lower and upper estimates.

[7] 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.

Concluding comment

The term impact evaluation has had many meanings in the past and continues to be used in various ways. The particular definition used here – counterfactual analysis of final welfare outcomes – is one which implies the need for a rigorous approach to establishing the counterfactual. Two problems arise when attempting to do this – contamination and selection bias. This chapter briefly mentioned how these problems may be tackled. The next chapter goes into different approaches to rigorous impact evaluation in more detail.

2. Approaches to Impact Evaluation

Introduction

The topic of approaches to impact evaluation can be divided up in various ways. Amongst practitioners of rigorous impact evaluation the discussion currently focuses on the merits and drawbacks of experimental approaches and the usefulness of the alternatives. But before getting to that discussion it is worth spending a moment considering non-statistically based approaches. IEG supports quantitative impact evaluation, but lessons can be drawn from these other approaches. Along with rigor, IEG's approach is characterized as strongly rooted in the context of the intervention in a way intended to produce policy relevant results. In practice this means adopting a theory-based evaluation design. An appreciation of how to do this is just as important to a good impact evaluation as an understanding of the appropriate statistical technique.

A theory-based approach to impact evaluation

A theory-based evaluation design is one in which the analysis is conducted along the length of the causal chain from inputs to impacts. Many impact evaluations concern themselves only with the final link in the chain: final outcomes. But to do this is often to lose the opportunity to learn valuable policy lessons about why an intervention has worked (or not), or which bits have worked better than others.

Applying a theory based approach requires mapping out the channels through which the inputs are expected to achieve the intended outcomes. In many cases this analysis will already be contained in the project log frame. The log frame (see Box 2.1) may also specify indicators at the various levels. Indeed the M&E system may have collected these indicators for project areas, and can be a useful source of analysis of process aspects of the intervention.

A theory-based approach examines the links in the causal chain. Were there missing or weak links? There can be missing links if the project design missed some key determinants at the next level it should have sought to influence. For example, the nutrition project in Bangladesh attempted to change child feeding practices by providing nutritional counseling to mothers, but there are other actors who play at least as an important part in the decision making process, so the failure to include them constituted a missing link (Chapter 6). Social funds make assumptions about the nature of the community and community participation which are frequently not valid (see below), so that the selected projects may not be the priority for a large proportion of the beneficiary population (World Bank, 2002). There may also be weak links in the chain as a result of poor implementation.

Box 2.1 The Logical Framework (log frame)

The logical framework was designed as a project planning tool, describing how it is that the inputs will achieve the desired objectives. The terms used vary by agency, but broadly correspond to inputs, activities, outputs, and outcomes. Indicators are defined as part of the logframe matrix which measure performance in delivering inputs, carrying out activities, producing outputs and achieving outcomes.

The logical framework used by the World Bank is a dynamic 4 X 4 matrix. The first column represents the project goal, objective, output (or deliverables) and activities. The second column specifies how to measure whether the activities, outputs and outcomes of the project have been carried out and accomplished. It contains the monitoring and evaluation indicators, including the targets. In the third column the sources of information for the indicators are identified. This column describes the monitoring and evaluation system. Finally, the fourth column lists the assumptions which are critical for the success of the project, but which are out of the project's control.

Project summary	Performance indicators	Monitoring & Evaluation system	Important assumptions
Goal The higher overall goal to which this project (together with other programs) will contribute.		Evaluation system	
Project objective The specific impact of the project: changes in welfare outcomes, or improvements in institutional performance. (The intended / assumed effect of the project outputs).	*Key performance indicators* - A few indicators that measure whether the project outputs have had the intended impact on children and other beneficiaries.	Evaluation system	Assumptions on the relationship between project impact and the overall goal.
Project outputs The project intervention: The outputs and deliverables that the project (team) are held accountable for.	*Output indicators* - To measure the value added of the project.	Monitoring system	Assumptions on the relationship between accomplished project outputs and project impact.
Project activities The specific activities that have to be carried out in order to accomplish each respective project output.	*Input indicators* - Usually the financial, physical and human resources needed to carry out the activities.	Monitoring system	Assumptions on the relationship between implemented project activities and outputs.

Of dummies and black boxes

The most usual approach to measuring impact is to examine the mean value of the indicator of interest in project and control areas, attributing the difference to the project. The regression based approach can give an identical estimate to the single or double difference approaches (see Box 2.2 and Appendix I) by use of a project dummy; that is a zero-one variable taking the value of 1 for observations in the project area. However this approach tells us nothing about the channels through which the project has its impact, and so is in general not consistent with the general philosophy of the theory-based approach.

Regressions can help open the black box through a two, or more, stage process. Impact can be examined by modeling the determinants of the outcome of interest. This approach can draw on well-established literatures modeling those outcomes.[8] Amongst the explanatory variables will be factors which are affected by the intervention. Knowing how much an explanatory variable has changed as a result of the intervention (which may be determined by examining double differences for that indicator, or itself be subject to regression analysis) allows calculation of the project effect via that particular channel. This is the approach taken in the most recent IEG studies. For example, regressions show the impact on improvements in the physical quality of schools on enrolments, and of teaching materials and methods on learning outcomes in Ghana (see Chapter 4). The project may have its effect not only by changing the quantity of a factor but its productivity. Regression analysis can readily accommodate this possibility by allowing the coefficient on the variable to vary before and after the intervention. The contribution of the changes in quantity and productivity can be calculated separately.[9]

When using the regression-based approach two points need be borne in mind. First, there have been important developments in micro-level statistical analysis in recent years, so the evaluator need be familiar with 'the state of the art' in their particular area in order to retain credibility. Second, the regression based approach may not remove the problem of selection bias. But it can do so if selection is based on observables, provided data are available on these observables and the selection process is correctly modeled. Hence it is very important that the evaluator be aware of who benefits from the intervention. There are a variety of statistical approaches which can be used to eliminate the selectivity bias.[10]

---

[8] For example the IEG studies of education in Ghana and health in Bangladesh began their work program by reviewing literature on determinants of school attainment and under-five mortality respectively (in economics terminology, the literature education and health production functions).

[9] This technique is known as the Oaxaca decomposition. An example of the approach is the Bangladesh health study carried out by the IEG, which showed how training traditional birth attendants had increased their effectiveness in reducing neo-natal mortality.

[10] For example, the IEG study of health in Bangladesh (Chapter 5) used a trivariate probit analysis of mortality to allow for the potential endogeneity of both using a trained birth attendant and immunization. This analysis thus also allowed discussion of which children are not being reached by the immunization program.

This is not so say that double difference estimates have no part to play in impact evaluations. In some cases project dummies may be the most appropriate way of capturing project impact. But these estimates should be situated in the context of an overall theory-based approach.

Box 2.2 Difference and double difference estimates

The difference between the outcome in the treatment group (project area) and comparison group is a single difference estimate. The validity of this estimate as an estimate of project impact requires that the treatment and control groups had the same values of the outcome prior to the intervention. If this is not so then the single difference estimate will be biased. If the treatment group already had superior outcomes prior to the intervention then their better performance post-intervention cannot all be attributed to that intervention. The double difference – which is the difference in the change, or, equivalently, the change in the difference – allows for this possibility. Double differencing removes time invariant differences in factors influencing the outcome between the project and comparison groups. However, the validity of the double difference estimate still relies upon the assumption that external determinants of the outcome were the same for treatment and comparison groups during the course of the intervention. For example, for an agricultural project both groups (areas) should have experienced similar rainfall patterns. Where these factors have varied then a regression based approach can control for these differences if they are observed. Hence data collection need cover not only the indicators directly relevant to the project, but also other determinants of the outcomes (and intermediate outcomes) of interest.

The attractiveness of double differencing is a main reason for the desirability of baseline data. Another advantage is that information on the pre-intervention characteristics of beneficiaries allows analysis of targeting and addressing the simple factual question of changes in outcomes amongst beneficiaries. In addition to baseline and endline surveys, a midterm survey is advantageous. Such a survey will allow an initial check on impact, and also allow analysis of changes in impact over time.

The importance of the factual

While a well-constructed counterfactual is central to establishing impact, the importance of the factual – what actually happened – should not be overlooked. Constructing a picture of how the intervention has played out on the ground, which nearly always requires data from the treatment group alone and not from a comparison group, is essential to a good impact evaluation, shedding light on the findings.

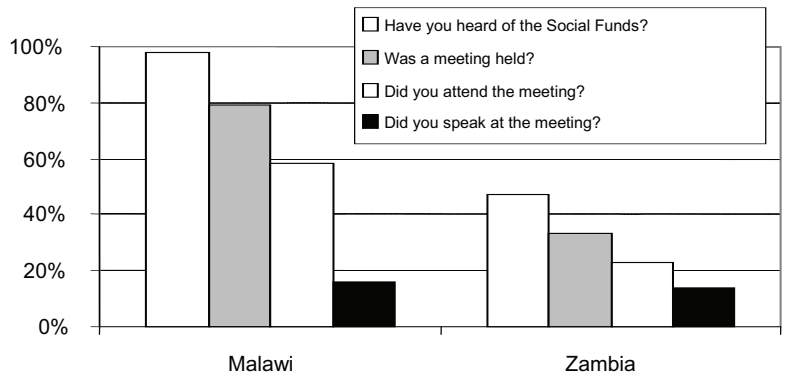
The IEG study of agricultural extension in Kenya, discussed in Chapter 7, found that there had been little change in extension practices during the lifetime of the Bank projects to support extension services, and that only a small minority (7 percent) of farmers participating in the project met extension agents as frequently as intended. Armed with this information, it comes as no surprise that it is difficult to discern any project impact. The evaluation of the Bangladesh Integrated Nutrition Project (Chapter 6) found that the majority of women and children eligible for supplementary feeding were not receiving it, thus undermining the potential beneficial impact of the intervention on community-level nutritional status.

An important example of the analysis of the factual is targeting – who participated in the project? The characteristics of those who participated should be compared to the general population, which may well require a different data set to that from the comparison group. Provided the survey data were collected in a comparable way, published secondary sources, such as the census or a national household survey, will suffice for these comparisons.

Constructing the factual often requires a combination of qualitative and quantitative information. An example of this fact is IEG's analysis of how social funds have operated at village level in Malawi and Zambia (World Bank, 2002).

Field visits in both countries revealed a picture in which a small number of people are instrumental in initiating the project and carrying it forward. To play this role needs someone not only with knowledge of the social fund, but the social and other skills (good literacy and numeracy) to carry forward the application – in the words of one headmaster in Zambia "someone who is not afraid to enter offices". This person is not the average villager, but more likely to be one of the few professionals in the community, such as teachers and health workers. However, as an outsider, these people are not in a position to mobilize the community, and it is here that traditional social structures come in to play. The headmaster may work through the PTA or sometimes directly with the headman. Following a decision by a small group to apply for social fund support for the school the PTA in Zambia will then seek the agreement of the village headmen, whereas in Malawi traditional leaders will mobilize the PTA. Hence identification of a particular sub-project usually takes place before the community becomes involved. IEG's quantitative survey data confirm the limited role of the meeting in project selection. Of all the respondents interviewed in the five communities in Zambia, only 33 percent knew of the meeting held for the sub-project selection, 23 percent attended and only 14 percent spoke at it (Figure 2.1). In Malawi far more knew of the meeting and attended it - 79 and 58 percent respectively – but only 16 percent spoke at it.

Figure 2.1 Participation in Subproject Selection



The headmen in turn seek the backing of the chief and then call the community together. At this meeting the community is told of the plan to apply for help rehabilitating the school, and that they need to provide labor. Since the community meeting with social fund staff or local officials takes place once considerable work has already been done there is little room at that stage for dissension even if the dynamic of a public meeting permits it. All households are expected to contribute and the headmen keep a register. Fines are also imposed on those who do not contribute, such as additional workdays or more arduous labor on the chief's land, though the fine may be money or livestock such as a chicken.

In summary, the majority of the community does participate actively in making bricks but rather more passively in decision making: something which is best understood from the qualitative data, but confirmed in a systematic way using the quantitative data. This factual analysis of the way in which projects are identified and implemented stands in contrast to the 'participatory model' put forward by many proponents of social funds. An understanding of this factual casts a great deal of light on why such projects have a limited impact on 'building social capital', since communities require a degree of social organization in order to apply for social fund resources.

Presenting impact: cost effectiveness and cost benefit ratios

The current focus on impact has been in some cases a backward step compared to earlier approaches which focused on cost effectiveness, or a full cost-benefit analysis with a rate of return. The finding that a project has a significant impact on school enrolments, nutritional status or firm profits is of little policy relevance if data are not presented on the cost of achieving that impact. At the very least, data on the cost per unit of output or outcome (e.g. how many dollars to get a child into school) should be presented as part of the impact analysis. In very many cases, impact evaluation can feed into a full *ex post* cost-benefit analysis (CBA). There are clear advantages to taking such approach, since CBA allows for other aspects of evaluation – such as assigning poverty weights, or valuing things not valued by the market – to be taken into account, which is not done in a straightforward impact analysis.

Approaches to constructing a comparison group[11]

Impact evaluation requires collecting data from both those affected by the intervention (the treatment group) and a similar group who have not been treated (the comparison group). This may be done by:

- Randomization
- Pipeline
- Matching areas on observables
- Propensity score matching

[11] More comprehensive treatment of the issues discussed here can be found in Baker (2000) and, more technically, Ravaillon (1999). An extended discussion of quasi-experimental designs can be found in Shadish *et al.* (2006).

Alternatively, differences in characteristics may be controlled for using a regression-based approach. These approaches are discussed in turn.

Experimental design: the random allocation of treatment[12]

Experimental design requires that the eligible population be identified and then a random sample of that population be 'treated', i.e. included in the project. For example, only 200 schools are chosen at random to be included in the project out of the 1,200 schools in 10 project districts. The untreated (or a random sample of the untreated) are a valid comparison group since there should be no systematic difference between their characteristics and those of the treatment group. It is in this case that the comparison group can properly go by the name of the control group, since the experimental approach implies that the evaluator controls the environment to ensure that the control does not become contaminated.

There are misconceptions about the randomized approach, so that it is held to be wholly inappropriate in a development setting. This is not so, and it has been successfully applied in several cases. Indeed, several of the claimed problems of a randomized approach are common to all impact evaluations. First, randomization is no more expensive than any other survey-based impact evaluation. Second, experimental design requires that beneficiaries are chosen at random from the eligible population, e.g. slum residents; there is no requirement at all that the population as a whole be considered for treatment. In the case of the school improvement project mentioned in the previous paragraph, a measure of targeting can still be achieved by selecting poor districts as the project districts. Third, allocating benefits to only a subset of potential beneficiaries is a result of the project budget constraint, not the decision to randomize. Hence there is nothing morally reprehensible about the decision to keep an untreated group – the same is true with any comparison group. Equally, the desire to keep an uncontaminated comparison is just as true as any impact study with a baseline.[13] Finally, a randomized design need not necessarily imply a black box approach, though this has indeed often been so in practice.

However, there <u>are</u> limits to the applicability of randomization in development evaluation. The first is that the evaluation design may perforce be ex post, so that the opportunity to randomize has long since passed. Second, the term 'treatment group' reflects the medical antecedents of the randomized approach. The medical analogy is apt since discrete, homogenous interventions – like taking a pill – are most amenable to a randomized approach. Where the nature of the intervention varies, then either multiple comparisons are required or an alternative needed which recognizes this heterogeneity.

---

[12] See the papers of Rawlings (2005), Kremer and Duflo (2005), and Ravallion (2005) – all published in an IEG conference volume *Evaluating Development Effectiveness* – for a more detailed discussion of randomization.

[13] Ravallion (2005) suggests that officials may be more inclined to offer non-participants 'compensatory programs' when they are not selected merely because of bad luck rather than some observable criteria, but does not offer examples of this problem in practice.

Many development interventions are complex in design, so that a randomized evaluation design may be appropriate for at best a subset of the intervention. Third, the experiment implies that the evaluator maintains control. This may not be possible. Those selected for the intervention may not want to take part, so selectivity bias comes back in. Or those not selected may lobby for inclusion, or for a comparable intervention, and so become contaminated. Or randomization may just prove to be a political non-starter. Other programs intend to be comprehensive in scope, such as attaining universal primary education. IEG studied the impact of debt relief under the Highly Indebted Poor Country (HIPC) initiative, under which all heavily indebted low-income countries, qualified for assistance, and so had to resort to more imaginative means of establishing a counterfactual. And projects working with a small number of entities, such as institutional development activities, cannot use a randomized approach.

Hence, experimental methods are in practice only applicable to a narrow range of the interventions supported by agencies such as the World Bank. Where they are applicable then they should be used, certainly more so than is done at present. Project managers need be made aware from the outset of the implications of randomization for program design. The evaluation design should incorporate study components of a qualitative nature and be sure to collect data across the log frame. Where experimental approaches are not applicable then the evaluator need turn to one of the alternatives discussed below.

Pipeline

The pipeline approach takes as the comparison group individuals, households or communities which have been selected to participate in the project, but not yet done so. In principle, there is therefore no selectivity bias, but this assumes that there has been no change in selection criteria, and that all applicants were not ranked and then the project 'worked down' the list. If the latter is the case then the approach ensures a bias rather than avoids it. Clearly the approach can only be used for activities which continue beyond the end of the project being evaluated.

IEG attempted an almost literal pipeline approach in its study of irrigation in Andhra Pradesh. The comparison group was selected from villages due to get water in subsequent seasons from the project-constructed canals. However, a check on comparability was made by estimating a propensity score using village-level census data. The region of common support turned out to be very small. Since the project was expanding the canal system by extending secondary and tertiary canals, the villages which were yet to be connected to the canal system were typically more remote than the already connected, with often quite different characteristics, most notably distance to nearest urban center.

Propensity score matching

Selection may be based on a set of characteristics rather than just one. Hence the comparison group need 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 (PSM) identifies a group of individuals, households or firms with the same observable characteristics as those participating in the project. It does this by estimating a statistical model of the probability of participating (propensity to participate) using a regression model with participation as the zero-one dependent variable, and a set of observable characteristics, which must be unaffected by the intervention, as the explanatory variables. The coefficients are used to calculate a propensity score, and participants matched with non-participants based on having similar propensity scores. In practice there are a range of ways of performing this matching, with the most common being to match each participant with their five 'nearest neighbor' non-participants (i.e. the five non-participants with the closest propensity score). The difference in the mean outcome from the two groups is taken as project impact.[14]

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. IEG's study of the Bangladesh Integrated Nutrition Project created a comparison group in this way using data from the Nutritional Surveillance Project, as the comparison data available had only a small sample size.

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)

The potential problem with PSM is that facing all quasi-experimental approaches: selection on unobservables. Unobservables which simply affect project outcomes and which are constant over time can be swept out by taking double difference estimates. But if they are time variant, or correlated with both selection and outcomes, then biased estimates will result.

[14] The theory underlying PSM is that matching on a linear combination of X characteristics in this way is an unbiased estimate of the result from matching individually on each of the X characteristics (something that would prove impossible to do in practice).

Regression-based approach

The regression based approach, modeling the determinants of outcomes and their intermediate values, was outlined in chapter 1. The approach has the advantage of flexibility – it does not lump different activities under the single heading of 'the intervention' – and automatically incorporates differing intensities of participation. It is only when the treatment is a simple, homogenous activity that dummy and mean comparison approaches are appropriate. However, the adoption of the regression-based approach does not mean that problems of selection bias are removed. They are not and must be addressed. Where selection is based on observables then this is readily done.

Better no numbers than silly numbers

Some interventions are not amenable to rigorous quantitative impact evaluation, for example when the universe treated is small, such as in an institutional development project with a single ministry. Even if the number of treated units is larger if the nature of the intervention, or the setting of the intervention varies, then quantitative analysis may not be useful. The BINP project discussed in Chapter 6 adopted a variety of designs (implemented by government or NGO, coverage of community workers, working intensively with newly married couples). As a result the design was different in each of the six initial project sub-districts from which data were collected for an evaluation of the project. This was intended as an experiment to see which approach worked best – but it was of course impossible to separate area effects from design effects.

When there are few observations to consider for each specific intervention being considered, then a survey-based quantitative approach is very unlikely to be fruitful. Careful qualitative analysis of institutional change and its drivers will be more revealing.

Data collection

Sources of survey data

Quantitative data for impact evaluation may come from four sources: (1) entirely from own survey, (2) piggy-backing, (3) synchronized survey, or (4) analyzing existing data sets.

The most usual source of quantitative data is to undertake a survey of both treatment and control areas. The clear advantage of this approach is that both the design of the survey instruments and the timing of the data collection can be tailored to meet the needs of the study, with the caveat that the evaluator is often brought into the picture too late to ensure a proper baseline. However, this is a costly option, and for that reason the sample size, especially for the control, is often less than would be desirable. It is also quite an undertaking to manage data collection to ensure data quality, especially in those countries in which the pool of skills in implementing household surveys is quite limited. In such cases, evaluation team members have to be closely involved in survey implementation, including training of enumerators and conducting the pilot. Some IEG studies have suffered from these problems as the data collected proved to be unusable owing to errors

made in the field. Poor management of data entry can also be a problem – in one IEG study one of the data managers "corrected" many questionnaires to accord with his own, incorrect, understanding of the questions. And in the same case, time was wasted working with poorly entered data before a decision was made to have the questionnaires sent to Washington to have the data re-entered.

Piggy-backing means joining forces with on-going survey effort, paying toward the survey costs so that data suitable for the evaluation are also collected. Piggy-backing can get round the quality problems just mentioned, since the large household survey which is being piggy-backed is most likely implemented by an agency well-skilled in undertaking such surveys, most usually the national statistics office. The data collection agency is requested to undertake two modifications to their survey design: (1) ensure adequate coverage of project areas by 'over-sampling' them (i.e. including more that would occur in the random sample design being applied), and (2) add a project-specific module. The project-specific module should cover process aspects of the project. But it should be designed in such a way that most, if not all, of the questions can be sensibly applied in non-project areas. Indeed this module is the ideal opportunity to collect data on similar interventions not falling under the project. Designing the module in this way will help reduce the bias introduced by asking leading questions, a problem which is addressed below.

An alternative way to utilize an on-going large scale survey is to use the larger survey to construct the control group, whilst collecting project area data through your own survey. This approach is called a synchronized survey as both design and timing have to be synchronized. Both surveys need have the same questions in order to allow matching on observable characteristics, which can be done using propensity score matching. The time of the surveys needs to be the same to control for seasonality in indicators. These are quite demanding requirements which have to be taken seriously. For its analysis of the Bangladesh Integrated Nutrition Project IEG used data from the monthly, nationally representative Nutritional Surveillance Survey to construct a control group using PSM.

The final alternative is to use existing data. Quality concerns are of course an issue, but there are many high quality surveys, such as Demographic and Health Surveys and income and expenditure surveys, which are in the public domain. This route is most cost effective. But it has the major drawback of the survey design not being oriented toward the intervention of interest. Data may be collected on some interventions (e.g. textbook supply or immunization), though 'process aspects' may be missing. For example, IEG's analysis of child health outcomes was able to document the link from the expansion of immunization coverage to reductions in under-five mortality. But the study could not look in detail at service delivery since there was insufficient detail on process indicators.

A variation to using existing data is to use an existing survey as a source of baseline data. Several IEG impact evaluations, such as that on agricultural extension in Kenya (Chapter 7), conducted a survey which repeated fieldwork in areas subject to a survey at the beginning of the project. The study of basic education in Ghana (Chapter 4), resurveyed

85 communities which had been subject to an income and expenditure survey, with an enhanced education module, fifteen years earlier, thus providing data on changes in educational performance and its determinants across a fifteen year interval.

The importance of good survey planning and implementation

The time required from initiating a small survey to going to the field should be at least three months, and six months is more realistic. A further two to three months need be allowed for data entry and initial cleaning. For a larger survey a longer lead time is necessary. Hence the time between initiating the survey and receiving the data can easily be one year or more, and this fact need be reflected in the evaluation timetable.[15] If piggy-backing an existing survey it is likely to be at least two years between initiation and receiving the data in a form ready for analysis.

Questionnaires are most usually designed by adapting an existing questionnaire to one's own purpose. There is nothing wrong with this approach, since it draws on existing experience. There are also manuals, most notably that for Living Standards Measurement Surveys (Glewwe and Grosh, 2000), which present and discuss questionnaire design. However, the survey instruments have to be adapted to the requirements of the evaluation. There are two aspects to this. The first is that the design should reflect the program theory you plan to test so that indicators are collected relating to the different links in the causal chain. A tabulation plan should be made showing the tables and analysis which will be carried out, which helps check that the necessary questions are asked and that unnecessary ones are not. It is important to avoid the temptation to include too much – most respondents will get respondent fatigue after an hour at most, so that the data collected after that point are of questionable value. The second is to adapt to local context, which is best done through qualitative information obtained through field exposure. Some data may be better collected at community or facility level. A health or school questionnaire is essential for social sector projects, and can be very powerful if the household questionnaire allows users to be linked to particular facilities.

The importance of pre-testing cannot be over-emphasized. Pre-testing should not be restricted to a pilot. Rather the questionnaire should be subject to a continuous process of review and revision from the time the first questions are written until the time it goes to the field. Members of the evaluation team should try out the questionnaire on each other, colleagues, friends and family. This should be done as seriously as possible, treating it as an actual interview. Ideally, the data from these tests should be entered and used to create tables according to the tabulation plan.

Recruiting good quality enumerators and making sure they are well trained is vital to the success of the evaluation. This is an activity in which core evaluation team members should be involved, it should not just be contracted out. Performing 'dry runs' of the questionnaire during the training is the most useful part of training to expose the

[15] Bamberger (2006) discusses various strategies for reducing these time-lines.

enumerators to the questionnaire and to continue testing the instrument. In addition to providing familiarity with the survey instruments, training should also provide enumerators with a general understanding of the study context, and also overcome biases which may affect their administration of the survey.[16] To avoid such biases, the enumerators should not have had any prior connection with the project. But the pilot is also important, and should be carried out amongst respondents similar to those who will be subject to the final questionnaire. All enumerators should participate in the pilot, to increase their exposure to the questionnaire and increase the number of pilot questionnaires administered. Once again, staff from the core evaluation team should take part in the pilot and it is very useful to keep at least one team member in the field throughout the survey. Not only will he or she act as an extra layer of supervision, they will gain valuable qualitative field experience.

Creating a baseline: the use of recall data

Where no baseline data were collected it may be tempting to generate baseline data by querying respondents about their pre-intervention circumstances. This approach can yield some information of value but need be used with caution. As with many aspects of survey design, a useful test is to ask whether you yourself could reliably answer such a question. Precise information on prices or production levels more than one year ago is not likely to be reliable. But remembering major events, including asset ownership, when changed employment, births and deaths are likely to be more reliably recalled, though even vital events have been shown to be subject to biases.

Particular care must be taken in establishing the reference period. Respondents are likely to 'telescope' and 'heap'. The former means that events are generally held to be more recent than they actually were. Heaping means that events (or ages) are placed at 5, 10 or 20 years, rather than more precise estimates. The most usual way to overcome these problems is to identify a reference event with which everyone is familiar at local or national level. Questions may then be asked that 'at the time of X did you….?'

Avoiding leading questions

A common error in survey design is to use leading questions. Many impact studies adopt a protocol by which the enumerator says, "I am looking at project X. Do you think project X was a good thing?" Such an approach will yield biased results, especially if the respondent hopes to receive future benefits from Project X. The ideal situation is one in which the respondent remains unaware of the subject of the evaluation, at least until late in the study. Indeed, the gold standard in medical research often requires that members of the research team themselves are unaware of the hypothesis under examination to avoid researcher bias. It is better that the enumerator introduce themselves as undertaking general research on health, agriculture or whatever.

[16] The problem of 'already knowing the answer bias' is not restricted to enumerators. In one IEG study the supervisor conducting the training discussed one question by saying "the answer to this one is no, so there's no need to ask, just put no".

Consistent with a theory-based approach, the questionnaire should not seek to ask directly if Project X affected outcome Y, but to measure indicators across the logical framework (see Box 2.1) of Project X. The design should allow for all the different factors which may have affected Y, including interventions supported by other agencies.

When such an approach is adopted then the questionnaire used in the treatment and comparison areas can be the same. When they are different then there is less basis for comparison. The end of the project questionnaire can include a project-specific module which can collect process data on the project. However, even some process data can be collected in a general questionnaire which asks about involvement in different activities, so the project focus is hidden. IEG is currently studying a project in Andhra Pradesh which encouraged the formation of self-help groups at village level. But the household questionnaire asks about participation in all groups. The identification as to which groups are supported by the project is made at the village level. Hence the nature of participatory processes can be compared between groups supported by the project and those which were not.

Combining quantitative and qualitative data collection

Good evaluations are almost invariably mixed method evaluations. Qualitative information informs both the design and interpretation of quantitative data. In a theory-based approach, qualitative data provide vital context, as in the example of social funds given above.

Many evaluations under-exploit qualitative methods, both in the techniques they use and the way in which analysis is undertaken. It is all too common-place to restrict data collection to key informant interviews and perhaps a few focus groups. But there are a far greater range of qualitative data collection methods, which can often produce more robust findings than can quantitative methods. Having said that, field experience by members of the core evaluation team (i.e. the people responsible for design and writing the final report) is an invaluable source of qualitative data which cannot be overlooked for good quality evaluations. And field experience means literally the field, not only meetings with government and project officials. It is very desirable to get such exposure very early on in the study so it can help inform the evaluation design. Later exposure will also help with interpreting the findings.

Concluding comment

Impact evaluation is approached in various ways. The current emphasis on rigor is important but should not lead to the neglect of aspects of evaluation design which ensure its policy relevance. Adopting a theory-based approach is the best way of ensuring such relevance, since it will yield information on how the program is working not just if it is working. Application of such an approach implies a mixed-methods evaluation design, i.e. one that combines quantitative and qualitative data collection. The time required for proper survey design should not be under-estimated, and can provide an opportunity for the evaluation team to spend time in the field which will prove invaluable when it comes to writing a well-contextualized report.

3. Impact Evaluation in IEG

The Role of IEG

IEG is an independent unit within the World Bank, and it reports directly to the Bank's Board of Executive Directors. IEG assesses what works, and what does not; how a borrower plans to run and maintain a project; and the lasting contribution of the Bank to a country's overall development. The goals of evaluation are to learn from experience, to provide an objective basis for assessing the results of the Bank's work, and to provide accountability in the achievement of its objectives. It also improves Bank work by identifying and disseminating the lessons learned from experience and by framing recommendations drawn from evaluation findings.

The World Bank's operational staff conducts separate, self-evaluations of the Bank's projects as soon as the project is completed; currently, about 270 new projects are completed each year. These self-evaluations are then assessed independently by IEG staff, using a rapid review approach based on standardized criteria.[17] IEG's evaluation approach is termed objectives-based evaluation, focusing on whether a project's actual outcomes are likely to achieve its stated objectives. The specific criteria include:

- The *relevance* of the project's objectives in relation to country needs and institutional priorities;
- Its *efficacy* – the extent to which its development objectives have been (or are expected to be) achieved;
- Its *efficiency* – the extent to which its objectives have been (or are expected to be) achieved without using more resources than necessary;
- The *sustainability* of the project – the likelihood that its estimated net benefits will be maintained or exceeded over the life of the project;
- The *institutional development impact* – the extent to which the project improves the ability of a country to make better use of its resources; and
- The *performance of the Bank and the borrower*, focusing on how good a job each partner has done at each stage of the project cycle.

IEG's project ratings may well diverge from those of Bank operational staff. IEG also subjects 25% of completed Bank projects to detailed field inspections. These project ratings are then used as a building block for the conduct of IEG's higher-level evaluations – these comprise sector and thematic evaluations, country assistance evaluations, and evaluations of global programs.

Evaluation elsewhere in the World Bank

As in other development agencies, the evaluation work undertaken by IEG is only a fraction of all evaluation supported by the Bank. In the case of the World Bank, impact

[17] These are summarized in http://www.worldbank.org/ieg/IEG_approach_summary.html.

evaluation is also undertaken by borrower governments of Bank-supported projects or by the Bank's operational departments. These studies are called self-evaluations, as opposed to the independent evaluations undertaken by IEG. The Bank's research department (DEC) also undertakes evaluation work, usually impact evaluations. Unlike IEG, DEC's mandate does not require it to restrict its attention to activities supported by the Bank, so some studies are of less direct relevance in assessing the Bank's development effectiveness. However, DEC has launched an important new initiative – Development Impact Evaluation (DIME) – which provides technical support for operational staff wanting to include impact studies in their project designs, and hosts seminars and workshops to disseminate methods and findings from impact evaluations.[18]

IEG's approach to impact evaluation

As is clear from the above, impact evaluations are just one of a range of evaluation products produced by IEG. The Department also carries out Country Assistance Evaluations, sector studies, and project-specific studies called Project Performance Assessment Reports (PPARs) for 25 percent of all completed Bank projects. All these studies address the question of impact. However, there are also separate larger studies focused more specifically on impact. IEG has been engaged in conducting such impact evaluations over the past 20 years, producing on average around one report a year which attempts to look at project impact using a counterfactual.[19] Under its current program IEG is committed to producing a new impact study each year, and hopes to increase this number. The most recent of these studies are reviewed in the following chapters, with shorter summaries of others given in appendix II.

In addition to independence and its mandate to focus on Bank operations, IEG also brings a particular approach to undertaking impact evaluation.[20] This approach, stressed in this brochure, is to seek rigor in determining impact, but to do so within a well contextualized framework. This contextualization is based on adopting a theory-based approach which seeks to document the causal chain from inputs to impact. The advantages of this approach are to establish an argument of plausible association where circumstances preclude more formal establishment of impact and to ensure policy relevance of the study.

[18] The DEC website http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,menuPK:384336~pagePK:149018~piPK:149093~theSitePK:384329,00.html contains many useful resources for impact evaluation.

[19] A larger number of reports than this are classified as Impact Evaluation Reports, reflecting the different uses of the term impact (see chapter 1 above). However, the discussion here is restricted to those studies employing a counterfactual. The IEG Working Paper by Gupta Kapoor (2002) provides a full list and discussion.

[20] This brochure lays out this approach in more detail. Other statements, and IE reports, can be downloaded from www.worldbank.org/ieg/ie.

The timing of evaluation

To date all IEG impact evaluations have been undertaken ex post. Indeed, the majority take place some years after the intervention has been completed. The rationale for waiting some time can sometimes be that impacts take a long time to appear,[21] but is more usually to see whether there has been a lasting impact, i.e. any sign of project results once the project itself has packed up. There is much to be said for collecting the endline data not immediately on project closing but a few years later. There are however two problems. First, after some years poor institutional memory can mean that it is difficult to trace what the project actually did in different areas – a complaint for example in IEG's study of the Kampung Improvement Program in Indonesia for which data were collected five years after the project closed. Second, for evaluations financed by the project there is a practical constraint that the evaluation then the project must, by definition, still be open when the evaluation takes place, thus constraining the choice of timing. But an independent evaluation office, such as IEG, suffers from no such constraint.

But there are drawbacks in *ex post* evaluation designs, with greater ingenuity required to overcome selection problems. Prospective evaluation is intended to overcome this problem by ensuring that a sound evaluation design is in place from the outset. Prospective simply means that the evaluation team is involved in evaluation design at the outset of the project, to the extent that this may affect aspects of program design (e.g. randomization). Of course the independence of the evaluator should not be compromised, so that those responsible for program design and evaluation design need a degree of separation. Prospective evaluation means putting in place a design now from which evaluation results will only be available in 5-7 years.[22] But the wait should ensure good quality evaluation results. IEG is working to introduce prospective evaluation into its work program. The DIME initiative of the Bank's research department is supporting prospective evaluations in new Bank operations.

The cost of impact evaluation

In IEG's experience, rigorous impact evaluation is expensive, costing between US$ 300,000 and US$ 500,000 per study.[23] Costs might be lower for a very project-specific study, but are likely to be of this order of magnitude if data collection is involved. This high cost is one reason why impact evaluation, whilst an important part of the evaluation toolkit, cannot reasonably be expected to be carried out for any but a relatively small percentage of total operations. It is therefore important to select subjects carefully so as to build up a collection of policy-relevant knowledge.

[21] Impact evaluation is sometimes equated with allowing a time lag before undertaking the study under the mistaken belief that impacts necessarily take time to appear. But this is not necessarily so. For example, a feeding program for pregnant women to reduce low birth weight must, by definition, have an impact in less than nine months. Other impacts, such as behavior change or increased agricultural value added from feeder roads may indeed take longer to be realized.

[22] There is a mistaken belief that ex post evaluations are less relevant as they produce results on past programs (i.e. they are out of date) whereas prospective evaluation gives results for current programs. But in fact prospective evaluation is simply planning to get 'out of date' information in seven years time!

[23] Bamberger (2006) discusses ways in which this cost might be brought down.

4. Case Study 1: Improving the Quantity and Quality of Basic Education in Ghana

Introduction

In 1986 the Government of Ghana embarked on an ambitious program of educational reform, shortening the length of pre-University education from 17 to 12 years, reducing subsidies at the secondary and tertiary levels, increasing the school day and taking steps to eliminate unqualified teachers from schools. These reforms were supported by four World Bank credits – the Education Sector Adjustment Credits I and II, Primary School Development Project and the Basic Education Sector Improvement Project. The IEG study set out to examine what had happened in basic education[24] during this period.

Data and methodology

In 1988/89 Ghana Statistical Service (GSS) undertook the second round of the Ghana Living Standards Survey (GLSS 2). Half of the 170 areas surveyed around the country were chosen at random to have an additional education module, which administered math and English tests to all those aged 9-55 years with at least three years of schooling and surveyed schools in the enumeration areas. Working with both GSS and the Ministry of Education, Youth and Sport (MOEYS), IEG resurveyed these same 85 communities and their schools in 2003, applying the same survey instruments as previously. In the interests of comparability, the same questions were kept, although additional ones were added pertaining to school management, as were two whole new questionnaires – a teacher questionnaire for five teachers at each school and a local language test in addition to the math and English tests. The study thus had a possibly unique data set – not only could children's test scores be linked to both household and school characteristics, but this could be done in a panel of communities over a fifteen year period. The test scores are directly comparable since exactly the same tests were used in 2003 as had been applied fifteen years earlier.

There was no clearly defined 'project' for this study. The four projects had supported a range of activities, from rehabilitating school buildings to assisting in the formation of community-based school management committees. To identify the impact of these various activities a regression-based approach was adopted which analyzed the determinants of school attainment (years of schooling) and achievement (learning outcomes, i.e. test scores). For some of these determinants – notably books and buildings – the contribution of the World Bank to better learning outcomes could then be quantified. The methodology thus adopted a theory-based approach to identify the channels through which a diverse range of interventions were having their impact.

[24] Basic education comprises primary (Grades 1-6) and Junior Secondary (Grades 7-9).

As discussed below, the qualitative context of the political economy of education reform in Ghana at the time proved to be a vital piece of the story.

Findings

The first major finding from the study was the factual. Contrary to official statistics, enrolments in basic education have been rising steadily over the period.[25] More strikingly still, learning outcomes have improved markedly: 15 years ago nearly two-thirds (63 percent) of those who had completed grades 3-6 were, using the English test as a guide, illiterate. By 2003 this figure had fallen to 19 percent.[26]

Also striking are the improvements in school quality revealed by the school-level data: For example:

- In 1988, less than half of schools could use all their classrooms when it was raining, but in 2003 over two-thirds can do so.
- Fifteen years ago over two-thirds of primary schools reported occasional shortages of chalk, only one in 20 do so today, with 86 percent saying there is always enough.
- The percentage of primary schools having at least one English textbook per pupil has risen from 21 percent in 1988 to 72 percent today and for math books in Junior Secondary School (JSS) these figures are 13 and 71 percent, respectively.

School quality has improved across the country, in poor and non-poor communities alike. But there is a growing disparity within the public school sector. Increased reliance on community and district financing has meant that schools in relatively prosperous areas continue to enjoy better facilities than do those in less well off communities.

The IEG study argues that Ghana has been a case of a quality-led quantity expansion in basic education. The education system was in crisis in the seventies; school quality was declining and absolute enrolments falling. But by 2000, over 90 percent of Ghanaians aged 15 and above had attended school compared to 75 percent 20 years earlier. In addition, drop-out rates have fallen, so completion rates have risen: by 2003, 92 percent of those entering grade 1 complete Junior Secondary School (grade 9). Gender disparities have been virtually eliminated in basic enrolments. Primary enrolments have risen in both disadvantaged areas and amongst the lowest income groups. The differential between both the poorest areas and other parts of the country, and between enrolments of the poor and non-poor, have been narrowed but are still present.

[25] The discrepancy is readily explained and illustrates the superiority of survey data in tracking enrolment trends (see Annex H of IEG's impact study of Ghana basic education, World Bank, 2004).
[26] The finding of improved learning outcomes flies in the face of qualitative data from many, though not all, 'key informant' interviews. But such key informants display the middle class bias which persists against the reforms which were essentially populist in nature.

Statistical analysis of the survey results showed the importance of building school infrastructure on enrolments. Building a school, and so reducing children's travel time, has a major impact on enrolments. While the majority of children live within 20 minutes of school, some 20 percent do not and school building has increased enrolments among these groups. In one area surveyed, average travel time to the nearest school was cut from nearly an hour to less than 15 minutes with enrolments increasing from 10 to 80 percent. In two other areas average travel time was reduced by nearly 30 minutes and enrolments increased by over 20 percent. Rehabilitating classrooms so that they can be used when it is raining also positively affects enrolments. Complete rehabilitation can increase enrolments by as much as one third. Across the country as a whole, the changes in infrastructure quantity and quality have accounted for a 4 percent increase in enrolments between 1988 and 2003, about one third of the increase over that period. The World Bank has been the main source of finance for these improvements. Before the first World Bank program communities were responsible for building their own schools. The resulting structures collapsed after a few years. The Bank has financed 8,000 school pavilions around the country, providing more permanent structures for the school which can better withstand the weather.

Learning outcomes depend significantly on school quality, including textbook supply. Bank-financed textbook provision accounts for around one quarter of the observed improvement in test scores.[27] But other major school-level determinants of achievement such as teaching methods and supervision of teachers by the head teacher and circuit supervisor have not been affected by the Bank's interventions. The Bank has not been heavily involved in teacher training and plans to extend in-service training have not been realized. Support to "hardware" has been shown to have made a substantial positive contribution to both attainment and achievement. But when satisfactory levels of inputs are reached — which is still far from the case for the many relatively deprived schools — future improvements could come from focusing on what happens in the classroom. However, the Bank's one main effort to change incentives — providing head teacher housing under the Primary School Development Project in return for the head teacher signing a contract on school management practices — was not a great success. Others, notably DFID and USAID, have made better progress in this direction but with limited coverage.

The policy context, meaning government commitment, was an important factor in making the Bank's contributions work. The government was committed to improving the quality of live in rural areas, through the provision of roads, electricity and schools, as a way of building a political base. Hence there was a desire to make it work. Party loyalists were placed in key positions to keep the reform on track, the army used to distribute textbooks in support of the new curriculum in the early 1990s to make sure they reached schools on time, and efforts made to post teachers to new schools and make sure that they received their pay on time. Teachers also benefited from the large civil service salary increase in the run up to the 1992 election.

[27] The Bank projects financed the provision of 35 million math and English text books during the period under review.

Better education leads to better welfare outcomes. Existing studies on Ghana show how education reduces fertility and mortality. Analysis of IEG's survey data shows that education improves nutritional outcomes, with this effect being particularly strong for children of women living in poorer households. Regression analysis shows there is no economic return to primary and JSS education (i.e. average earnings are not higher to children who have attended primary and JSS compared to children who have not), but there is a return to cognitive achievement. Children who attain higher test scores as a result of attending school can expect to enjoy higher income; but children who learn little in school will not reap any economic benefit.

Some policy implications

The major policy finding from the study relates to the appropriate balance between hard and software in support for education. The latter is now stressed. But the study highlights the importance of hardware. In the many countries and regions in which educational facilities are inadequate then hardware provision is a necessary step in increasing enrolments and improving learning outcomes. The USAID project in Ghana encourages teachers to arrange children's desks in groups rather than rows – but many of the poorer schools don't have desks. In the words of one teacher, "I'd like to hang posters on my walls but I don't have posters. In fact, as you can see, I don't have any walls".

These same concerns underlie a second policy implication. Central government finances teacher's salaries and little else for basic education. Other resources come from donors, districts or the communities themselves. There is thus a real danger of poorer communities falling behind, as they lack both resources and the connections to access external resources. Hence children of poorer communities are left behind and account for the remaining illiterate primary graduates which should be a pressing policy concern.

The study highlighted other areas of concern. First amongst these is low teacher morale, manifested through increased absenteeism. Second is the growing importance of the private sector, which now accounts for 20 percent of primary enrolments compared to 5 percent 15 years earlier. This is a sector which has had limited government involvement and none from the Bank.

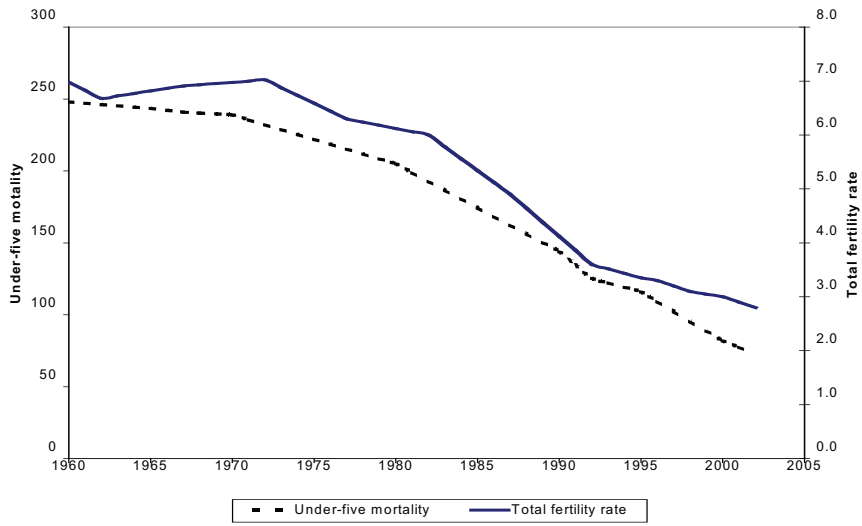
5. Case Study 2: Meeting the Health MDGs in Bangladesh

Introduction

Bangladesh began the 1970s as a new country in a dire situation. The ravages of war and famine meant that the prospects for development appeared bleak. Social indicators were amongst the worst in the world. Women could expect to have on average seven children during their child bearing years, but two of those would die before reaching their fifth birthday. Three-quarters of all children were malnourished.

Thirty years on the situation has changed drastically. The total fertility rate has fallen from seven to less than three, and under-five mortality from over 250 per 1,000 live births to around 80 by 2004 (Figure 5.1). These rates of progress mean that Bangladesh is on track to meet the Millennium Development Goals. Malnutrition remains high but has begun to decline in the last decade. IEG (2005) examined the factors underlying this success.

Figure 5.1 Both fertility and under-five mortality have fallen



Methodology

The IEG study utilized existing data sets. The analysis drew on both cross-country data, from a variety of sources, and national data mainly from the Demographic Health Surveys of 1992/93, 1996/97 and 1999/00. Multivariate analysis of the determinants of health and nutrition outcomes was carried out. This approached allowed the

identification of interventions in a range of sectors which had affected health outcomes. Whilst it was possible to carry out cost-effectiveness analysis, a full theory-based approach could not be applied because of the absence of process indicators.

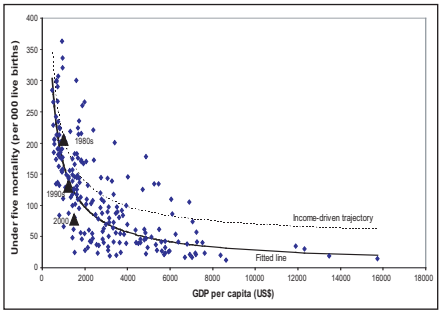
Findings

Economic growth is usually seen as a critical factor in reducing poverty in its various dimensions. Bangladesh is no exception to this point, and the country's respectable growth record has indeed played a part in the country's improved social outcomes. But it is not the whole story. Figure 5.2 shows under-five mortality and fertility plotted against income per capita for a cross-section of 78 countries at different points in time. Each data point represents the decade averages of income and the social outcome shown, using values from the 1970s to the current decade, so that there are up to four observations for each country.

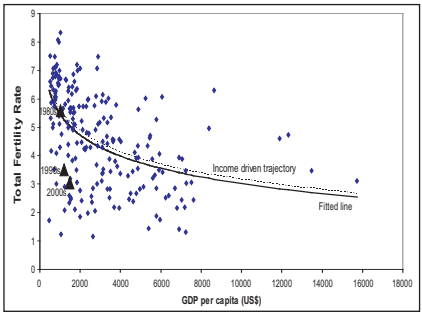
The solid line in each figure is the average relationship between income and the social outcome, that is "the fitted line". In the 1980s Bangladesh (indicated by the triangular data points, each labeled by its decade) lay above the average for under-five mortality and fertility, meaning that those indicators were worse than should be expected for a country at its income level. If these indicators had improved following the internationally-established relationship with income then subsequent observations for Bangladesh would have laid along the dashed line.[28] But in fact these later observations lie below the fitted line, showing that Bangladesh now does better than expected for a country at its income level. This finding suggests that there have been important, non-income-related, factors behind the improvement in mortality and fertility in Bangladesh.

Figure 5.2 Bangladesh's improvement in social outcomes is greater than can be explained by economic growth alone

(a) Under-five mortality



(b) Total fertility rate



Note: See discussion in text for explanation.

[28] The dashed line has the same slope as the fitted line, but with the intercept adjusted to pass through the 1980s observation for Bangladesh. There is an assumption here that the outcome-income relationship is the same for different countries over time. This is clearly not so, most notably in the case of the Bangladesh. The point of the analysis is to show quite how much Bangladesh departs from international norms so as to pose the question as to why this is.

Table 5.1 Growth in GNP Per Capita Accounts for at Most One-Third of The Reduction in Mortality… and Less Than a Fifth of Lower Fertility

	1980 actual	*2000 actual*	*2000 income-based estimate*	*Percent reduction explained by income*
Under-five mortality	205.0	77.5	163.1	32.9
Total fertility rate	5.6	3.0	5.2	16.0

Source: calculated from data used for Figure 5.1

The numbers behind these graphs provide an upper estimate of the extent to which growth in GDP per capita has contributed to improved social outcomes in Bangladesh (Table 5.1).[29] For example, under-five mortality was 205 per 1,000 live births in the 1980s. Income growth alone would have reduced it to 163 by 2000, but by then the actual rate was 78. Hence at most just under one-third of the improvement comes from higher average income. For fertility the share of income is even less, explaining at most 16 percent of the observed reduction.

The question of what then explains the additional reduction was analyzed through multivariate analysis of both cross-country and household data. The results revealed the following regarding selected interventions:

- Immunization coverage was at less than 2 percent in the early 1980s, but grew in the latter part of the decade (largely with the support of UNICEF and WHO, but later also other donors including the World Bank) so that by 1990 close to half of all children were fully vaccinated in their first 12 months. Immunization has averted over 2 million child deaths in the last two decades, at a cost of between $100 and $300 per life saved.

- The World Bank financed the training of approximately 14,000 traditional birth attendants (TBAs) until the late nineties, at which point training TBAs was abandoned following a shift in international opinion toward a policy of all births being attended by Skilled Birth Attendants. However, the evidence presented in this report shows that training TBAs saved infant lives, at a cost of $220-800 per death averted.

- Female secondary schooling expanded rapidly in the 1990s, especially in rural areas partly as a result of the stipend paid to all female students in grades 6-10 in rural areas supported by Norwegian aid, the Asian Development Bank, the World Bank and government. Amongst the benefits of the increase in female secondary schooling are lower infant and child mortality, at a cost of $1,080-US$5,400 per death averted.

[29] It is an upper estimate, since the calculations are based on a simple regression. This equation is undoubtedly mis-specified resulting in omitted variable bias. Income is positively correlated with several determinants of these outcomes – such as education and immunization – so that there will be an upward bias on the estimated regression coefficient for income.

- Rural electrification, supported through three World Bank programs in the 1980s and 1990s, reduces mortality through income effects, improving health services, making water sterilization easier and improving access to health information, especially from TV. Taking these various channels into account means that children in households receiving electrification have an under-five mortality rate 25 per 1,000 lower than that of children in non-electrified households.

Policy implications

The IEG study had the following policy implications:

- Publicly-provided services, with external support, were an efficacious and cost-effective means of improving health outcomes.

- Interventions from several sectors improved health outcomes. But this multi-sectoral causation did not mean that interventions had to be delivered in a multi-sectoral manner.

- Local evidence needs to be taken into account in making resource allocation decisions. The training of TBAs was abandoned in Bangladesh following international fashion, but local evidence shows it to have been effective in reducing infant mortality.

6. Case Study 3: The Bangladesh Integrated Nutrition Project

Introduction

While rapid strides were being made in reducing mortality and fertility in Bangladesh in the 1980s, malnutrition showed no improvement, affecting close to seventy percent of all children under-five. In order to address this remaining problem the government undertook a pilot nutrition intervention supported by the World Bank, the Bangladesh Integrated Nutrition Project (BINP).

BINP had three components of which the main one – the Community-Based Nutrition Component (CNBC) – was the focus of the IEG study. In each project thana (sub-district) a number of Community Nutrition Promoters (CNPs) were recruited, these being local women with children of their own and having achieved at least an eighth grade education. The CNPs implemented activities at the community level: monthly growth monitoring for children under 24 months old, supplementary feeding for malnourished children and pregnant women, and nutritional counseling in a variety of settings. The CNPs were overseen by a Community Nutrition Officer, and supervisors at the thana level who were staff of the implementing NGO.[30]

Data from the project monitoring system showed that BINP was having wide coverage, indeed exceeding project targets. There were also early signs of success in reducing severe malnutrition, resulting in the decision to scale up the project under the National Nutrition Project (NNP). However, this decision resulted in some debate following a study by Save the Children UK suggesting that the project had had little impact on nutritional outcomes. The IEG study was thus undertaken in a fairly highly charged atmosphere, but also one open to lesson learning if the project needed to adapt.

Data and methodology

The IEG study had three separate sources of survey data: (1) the BINP project evaluation, which collected data at baseline, midterm, and endline, (2) the survey carried out by Save the Children UK at the end of the project (plus some data from the registers kept by community nutrition workers), and (3) the Nutritional Surveillance Project conducted by Helen Keller International (a monthly survey of nutritional status). Between them these data covered not only nutritional outcomes but also a wide range of process indicators, allowing the application of a theory-based approach. The analysis was also informed by a reading of the anthropological literature, notably regarding the status of women.

The links in the causal chain for both growth monitoring/nutritional counseling and

[30] In some thanas BINP was implemented by GoB, so that this last level of supervision was missing.

supplementary feeding were examined. The key assumption behind CNBC was that "bad practices" are responsible for malnutrition in Bangladesh. This point of view was strongly argued in the BINP appraisal document: "behaviors related to feeding of young children have at least as much (if not more) to do with the serious problem of malnutrition in Bangladesh as poverty and the resultant household food insecurity do" (BINP SAR: para 1.13, p.4; World Bank, 1995). Therefore changing bad practice to good will bring about nutritional improvements. There are a number of steps in the causal chain behind this approach:

- The right people (those making decisions regarding under-nourished children) are targeted with nutritional messages
- These people participate in project activities, and so are exposed to these messages
- Exposure leads to acquisition of the desired knowledge
- Acquisition of the knowledge leads to its adoption (i.e., a change in practice)
- The new practices make a substantial impact on nutritional outcomes

A feeding program for malnourished children and pregnant women was implemented alongside growth monitoring. For this program to work:

- The target groups have to enroll in the program
- The criteria are correctly applied in selecting those to receive supplementary feeding
- Those selected for supplementary feeding attend sessions to receive the food.
- There is no leakage (e.g., selling of food supplements), or substitution (reducing other food intake)
- The food is of sufficient quantity and quality to have a noticeable impact on nutritional status

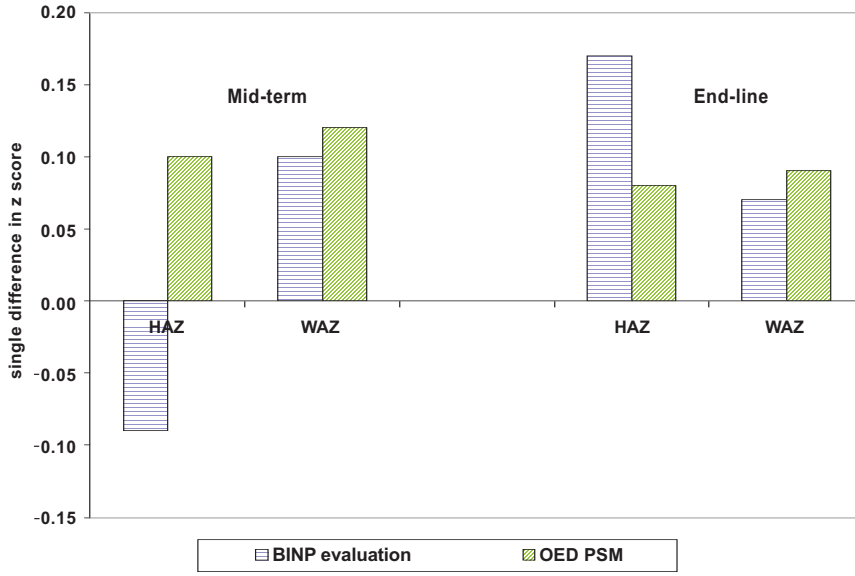
Findings

Save the Children argued that BINP had had no impact on nutritional outcomes. Another study by the Government of Bangladesh (Haider et al, 2004) also found little evidence of such an impact. Both these studies relied on single difference estimates at the end of the project, and supporters of BINP questioned the quality of the comparison areas selected. The evaluation commissioned by B INP itself showed some impact, but the quality of that comparison group was also questionable. Hence the IEG study compared the project area data from the BINP evaluation with a new comparison group constructed using propensity score matching from the Nutritional Surveillance Project data.

Figure 6.1 illustrates how the use of PSM improved the quality of the impact results. The figure shows single difference estimates of height for age (HAZ) and weight for age (WAZ) between project and control at both mid-term and endline. The BINP evaluation found WAZ to be better in project areas than the control, but height to be worse. By the endline it found both indicators to be better in the project areas, height more so than weight. Both these results are not intuitive. These indicators should move

in the same direction, though the impact on WAZ is likely to be greater than that on HAZ. This is exactly what the IEG analysis finds when using the NSP data to create a control group using PSM.

Figure 6.1 BINP impact on child nutrition using different control groups



The IEG study confirms that there is an impact from the project. But it is very low, especially at the endline, being equivalent to a reduction in malnutrition of less than 5 percent. This is not a very cost effective intervention.

Several weak and missing links in the causal chain are identified to explain this disappointing outcome. There were two missing links in the BINP chain. The first missing link was the relative neglect of some key decision makers regarding nutritional choices. Mothers are not the sole, or even main, decision makers for many factors (e.g. what to buy, as in rural areas men do the shopping) affecting child nutrition – husbands and mothers-in-law are also important, but were largely neglected in the delivery of nutritional messages. The second problem was the focus on pregnancy weight gain as pre-pregnancy nutritional status is the more important determinant of low birth weight. Even achieving the targeted improvements in pregnancy weight gain would have had only a small impact on the incidence of low birth weight.

There were also several weak links. Participation levels of the target audience were high, but many women escaped exposure to nutritional messages, and there was a high Type I error in the feeding programs. Critically for the nutritional counseling component, a substantial knowledge practice gap persisted, so many women did not put the advice they

received into practice, especially if they were resource or time constrained. In the qualitative fieldwork, women identified their mothers-in-law as a major constraint on adopting new practices. Those receiving supplementary feeding, often shared it with others or substituted it for their regular foodstuffs. This list of weak links in the chain explain why project impact was muted by the time final outcomes are considered. While attention can be paid to each of these weak links, the BINP experience does demonstrate the difficulty of implementing complex designs.

However, the project was not without success. It did mobilize the community around nutritional level, with high participation levels that had beneficial knock on effects, such as higher coverage of antenatal care in project areas. Positive nutritional outcomes were found for the most malnourished children, for whom it is possible that supplementary feeding really was supplementary.

Policy lessons

The IEG study confirmed the argument that scaling up the BINP model to the national level would be very costly and most likely not result in the desired improvements in nutritional outcomes. However, better targeting of both nutritional messages and supplementary feeding will improve project impact. But beyond that there is a need to examine more what can best be done to improve nutritional outcomes. These have improved across the country since the mid-90s, which the IEG report attributes to increased agricultural productivity mainly as a result of the adoption of high-yielding rice varieties. Hence the price of rice has fallen whilst incomes have grown.

7. Case Study 4: Agricultural Extension Services in Kenya

Introduction

The Training and Visit (T&V) system of agricultural extension sought to strengthen links between research and extension, and get these results to farmers through frequent visits by extension workers to farmers – at least monthly and often more frequently. The approach was widely adopted in the 1980s, with much World Bank support. However it began to attract critics. It was known to be costly and there was scant evidence of any impact on agricultural productivity. The IEG study (World Bank, 2000) entered this debate in order to provide such evidence for the case of Kenya.

Methodology

In 1997 IEG commissioned a survey of 285 households which had been covered by both the Rural Household Budget Survey (RHBS) in 1982, the year of the start of the first World Bank supported National Extension Project, and a Bank-sponsored survey in 1990. Data on these households were thus available at three points in time across a 15 year period, allowing a detailed picture to be built up of any changes in agricultural production and practices. IEG's own questionnaire included a contingent valuation module, which asked farmers about their willingness to pay for extension services. Qualitative fieldwork was also undertaken to analyze farmer perceptions.

A theory-based approach was adopted so that the study considered a set of intermediate indicators as well as the impact on final outcomes.

Findings

The study suggested that agricultural extension has a potentially important role to play: the data showed that farmers who were aware of improved practices usually put them into effect. This suggests that lack of knowledge was indeed a constraint. But the Bank's two National Extension Projects (NEP I and II) were not seen as having alleviated that constraint and so had had no discernible impact on production.[31]

The use of the theory-based approach revealed several places in which the causal chain had broken down, so that the interventions were not functioning as planned. First, for much of the project, there was little if any link from research to extension, so that there were no new messages for extension workers to take to farmers. Hence when workers did visit farmers on a monthly basis both sides viewed these meetings as rather repetitive. By the time NEP II was started the main messages being delivered by extension services, regarding maize production, had already been adopted by virtually all farmers, so there was no scope

[31] The study exploited the availability of panel data to conduct a fixed effects multivariate analysis.

for further benefits without some innovation in the information being conveyed.

But in fact less than one-tenth (7 percent) of contact farmers were visited on a monthly basis. This fact reflected the failure of the institutional development aspect of NEP I and II. Most extension workers continued to operate in the same way as they had before the project, hence the project can be expected to have had little impact on farming practices. The failure of ID was partly a failure of incentives but also a result of the changing fiscal climate. Government resources were shrinking so the extension service became increasingly reliant on donor finance, most of which went to pay salaries so there was little left over to support intensive delivery of extension services. This situation meant that the project was not sustainable. And although farmers were indeed willing to pay for extension services, it was nowhere near what they were actually costing under this system.

Policy messages

The report pointed clearly to changes which needed to be made. In its current form extension services were having little impact and so by definition were inefficient. However, impacts were discernible for poorer areas: extension services had played a role in allowing less productive areas to catch up. But delivery had been concentrated in the more productive areas where they had least impact! The IEG study thus recommended a re-targeting to poorer areas. It also pointed out that messages had to be better tailored to the needs of farmers, rather than a uniform package being delivered to all farmers. One way to achieve this was to become less top down and more responsive to farmer demands. Finally, the rate of innovation did not support such an intensive delivery system: a leaner extension service with broader coverage would be more efficient.

Appendix 1: Some Impact Algebra

Single and double difference project impact estimates

Single difference

The single difference effect of an intervention (β_d) on a particular outcome (y) is the difference in average outcomes between project and control areas at the end of the project (i.e. using endline data):

$$\beta_d = (\bar{y}_1 - \bar{y}_0) \quad (1)$$

where $\bar{y}_1$ and $\bar{y}_0$ are the observed means of the outcome in the project and the control areas respectively. The same expression can be written in the form:

$$y_{ip} = \alpha + \beta_d P + \eta \quad (2)$$

where y_{ip} is the outcomes observed for individual i, and the p sub-script denotes project, i.e. whether the individual lives in a project area (p=1, control: p=0), and P is a variable taking the value of one if the observation is in the project and 0 otherwise. When P=0, the expected value of y from equation 2 is simply α, that is

$$E(y_i | P = 0) = \alpha = \bar{y}_0 \quad (3)$$

which follows since regressing a variable on a constant produces the mean of that variable as the estimate of the constant. It further follows that when regressing the outcome on a constant and a project dummy then the coefficient on the dummy must be the difference between the mean outcome for the project and control areas, i.e. the definition of the single difference measure of impact:

$$E(y_i | P = 1) = \alpha + \beta_d = \bar{y}_1 \quad (4)$$

Substituting $\alpha=\bar{y}_0$ into equation (4) gives equation (1), showing the equivalence of the regression approach and calculating the single difference estimate to the 'manual' approach given by equation (1).

The advantage of the regression-based approach is that the model can be extended to include other variables which may also be affecting the outcome independently of the project intervention. To do this, include control variables (X). We can also include interaction terms between control and project variables, which allow project impact to vary according to beneficiary characteristics:

$$y_{ip} = \alpha + \beta d_p + \alpha_1 X + \alpha_2 X d_p + \eta \quad (5)$$

Double difference

The single difference approach assumes that the project and control groups had the same values of the outcome prior to the project. If this is not the case then the single difference over-estimates project impact if outcomes were already better in the project area (and under-estimates it if they were worse). If baseline data are available we know outcomes prior to the project and may calculate instead the double difference estimate. Double differencing also controls for constant area specific effects influencing the outcome.

The double difference impact (β_{dd}) of the project is given by the difference between the differences in mean outcomes for control and project areas reported at the end-line and the base-line.

$$\beta_{dd} = (\bar{y}_{11} - \bar{y}_{10}) - (\bar{y}_{01} - \bar{y}_{00}) \quad (6)$$

Where y_{tp} are the outcomes, $t=1$ at the end-line and $t=0$ at the base-line, $p=1$ if the household is in the project group and $p=0$ if the household is in the control group. The same expression can be written in regression form as:

$$y_{ipt} = \alpha_0 + \alpha_1 t + \alpha_2 P + \beta_{dd} P t + \eta \quad (7)$$

where d's are dummies for $p=1$, $t=1$ and p and $t=1$. β_{dd} is the project effect as defined in equation (6). Table 2 helps show why this is so.

Table 1 Expected outcome (y) by time and place

	t = 0	t = 1
p = 0	α_0	$\alpha_0 + \alpha_1$
p = 1	$\alpha_0 + \alpha_2$	$\alpha_0 + \alpha_1 + \alpha_2 + \beta_{dd}$

Substituting the values from Table 2 into equation (6) gives:

$$\beta_{dd} = (\alpha_0 + \alpha_1 + \alpha_2 + \beta_{dd} - \alpha_0 - \alpha_1) - (\alpha_0 + \alpha_2 - \alpha_0) = \beta_{dd} \quad (8)$$

If there is no difference between project and control, as should be the case, then α_2 should be zero. This may not be the case as some of the determinants of outcomes vary between project and control areas. These factors can be controlled for if data are available by including individual characteristics X simply by running:

$$y_{ipt} = \alpha_0 + \alpha_1 t + \alpha_2 P + \beta_{dd} P t + \alpha_3 X + \eta \quad (9)$$

In addition, we can include interaction terms in order to test, for example, that more educated households (ED=1) are more likely to understand nutrition education messages and to modify behaviors:

$$y_{ipt} = \alpha_0 + \alpha_1 t + \alpha_2 P + \beta_{dd} P t + \alpha_3 X + \alpha_4 ED + \alpha_5 ED t + \alpha_6 ED P + \alpha_7 ED P t + \eta \qquad (10)$$

In summary, regression analysis of endline data include a project dummy which represents the pure project effect. The analysis of the pooled data from baseline and endline includes a project dummy, a time dummy, and the interaction between time and project dummies. It is the coefficient on the interaction term which measures project impact, that is the differential change in the outcomes over time in project areas respect to control areas, i.e. the double difference. The coefficient on the time dummy captures the autonomous growth in outcomes common to both project and control areas. The coefficient on the project dummy indicates any difference in outcomes at baseline; this coefficient should be zero. If it is not zero then it is preferable to include in the model control variables which also affect the outcome (which is anyhow to be preferred).

Appendix 2: Overview of Selected IEG Impact Evaluations

AGRICULTURE AND RURAL DEVELOPMENT

Case Study 1: Pakistan

Scarp Transition Pilot Project

Report No. 16840, 1997

The projects: Irrigation in Pakistan suffers from the "twin menaces" of salinity and waterlogging. These problems have been tackled through Salinity Control and Reclamation Projects (SCARPs), financed in part by the Bank. Whilst technically successful, SCARP tubewells imposed an unsustainable burden on the government's budget. The project was to address this problem in areas with plentiful groundwater by closing public tubewells and subsidizing farmers to construct their own wells.

Methodology: IEG commissioned a survey in 1994 to create a panel from two earlier surveys undertaken in 1989 and 1990. The survey covered 391 farmers in project areas and 100 from comparison areas. Single and double differences of group means are reported.

Findings: The success of the project was that the public tubewells were closed without the public protests that had been expected. Coverage of private tubewells grew rapidly. However, private tubewells grew even more rapidly in the control area. This growth may be a case of contagion, though a demonstration effect. But it seems more likely that other factors (e.g. availability of cheaper tubewell technology) were behind the rapid diffusion of private water exploitation. Hence the project did not have any impact on agricultural productivity or incomes. It did however have a positive rate of return by virtue of the savings in government revenue.

Case study 2: Philippines

Second Rural Credit Program

Report No. 4557

The project: the Second Rural Credit Projects (SRCP) operated between 1969 and 1974 with a US$12.5 million loan from the World Bank. SRCP was the continuation of a pilot credit project started in 1965 and completed in 1969. As its successful predecessor, SRCP aimed at providing credit to small and medium rice and sugar farmers for the

purchase of farm machinery, power tillers, and irrigation equipment. Credits were to be channeled through 250 rural banks scattered around the country. An average financial contribution to the project of 10% was required from both rural banks and farmers. The SRCP was followed by a third loan of US$22.0 million from 1975-77, and by a fourth loan of US$36.5 million that was still in operation at the time of the evaluation (1983).

Methodology: the study uses data of a survey of 738 borrowers (nearly 20% of total project beneficiaries) from seven provinces of the country. Data were collected through household questionnaires on land, production, employment and measures of standard of living. In addition, 47 banks were surveyed in order to measure the impact on their profitability, liquidity, and solvency. The study uses before-after comparisons of means and ratios to assess the project impact on farmers. National level data are often used to validate the effects observed. Regarding the rural banks, the study compares measures of financial performance before and after the project taking advantage of the fact that the banks surveyed joined the project at different stages.

Findings: the mechanization of farming did not produce an expansion of holding sizes (though the effect of a contemporaneous land reform should be taken into account). Mechanization did not change cropping patterns, and most farmers were concentrating on a single crop at the time of the interviews. No change in cropping intensity was observed, but production and productivity were found to be higher at the end of the project. The project increased the demand for both family and hired labor. Farmers reported an increase in incomes and savings, and in several other welfare indicators, as a result of the project. Regarding the project impact on rural banks, the study observes an increase in the net income of the sample banks from 1969 to 1975, and a decline thereafter. Banks' liquidity and solvency position was negatively affected by poor collection and loan arrears.

Case Study 3: Sri Lanka

Kurunegala Rural Development Project, and Second Rural Development Project

Report No. 16418, 1997

The projects: The two projects were Integrated Rural Development Projects, an approach adopted in Sri Lanka in the late seventies. The US$ 34 million Kurunegala Rural Development Project (KRDP) was the first attempt at multisectoral planning for an entire district. The project focused on the agricultural sector, especially paddy and coconut production. Productive services (inputs supplies, extension services and credit were also provided). The US$ 40 million Second Rural Development Project (SRDP) expanded the approach to two other districts, Matale and Puttalam, with an additional focus on forestry and fisheries in Puttalam and export crops in Matale.

Methodology: Secondary data on production increases, inputs and yields in project and non-project areas within the same district are used to calculate incremental benefits.

These data are used to construct farm models for producers of different crops, and hence calculate the return to the project. Non-project areas in the same districts benefited from interventions from other donors, so the returns are noted to be under-estimates. A second approach uses the before-intervention figures as the counterfactual, thus attributing all increases to the project, which is recognized as producing over-estimates of project impact. Qualitative work was also undertaken, which included questions on changes in income and the sources of those changes.

Findings:

- Project targets for increased area and productivity were not met, but these were somewhat unrealistic. The project did contribute to production increases of the targeted crops, and so higher incomes for beneficiaries.

- ERRs are reported by crop for each of the three districts, and are as follows:[32] paddy 15, 19 and 5; coconut: 7 in both areas; export crops: 5, and productive components: 9, 11 and 7.

- Other aspects of the quality of life were not captured by counterfactual analysis. The beneficiary assessment pointed in particular to the benefits from improved roads. Beneficiaries complained about lack of consultation in irrigation rehabilitation meaning that users' concerns were not addresses.

HEALTH, NUTRITION AND POPULATION

Case Study 4: India

Tamil Nadu Integrated Nutrition Project

Report No. 13783-IN

The project: The Tamil Nadu Integrated Nutrition Project (TINP) operated between 1980 and 1989, with a credit of US$32 million from IDA. The overall objective of the project was to improve the nutritional and health status of pre-school children, pregnant women and nursing mothers. The intervention consisted of a package of services including: nutrition education, primary health care, supplementary feeding, administration of vitamin A, and periodic de-worming. The project was the first to employ Growth Monitoring and Promotion (GMP) on a large scale. The evaluation is concerned with the impact of the project on nutritional status of children.

Methodology: The study uses three cross-sectional rounds of data collected by the TINP Monitoring Office. Child and household characteristics of children participating in the program were collected in 1982, 1986, and 1990, each round consisting of between 1000

[32] The ERRs reported here are a simple average of the estimates from the two methods employed in the study.

and 1500 observations. The study uses before-after comparisons of means, regression analysis, and charts to provide evidence of the following: frequency of project participation, improvement in nutritional status of participating children over time, differential participation and differential project impact across social groups. Data on the change in nutritional status in project areas are compared to secondary data on the nutritional status of children outside the project areas. With some assumptions, the use of secondary data, make the findings plausible.

Findings: The study concludes that the implementation of Growth Monitoring and Promotion programs on a large scale is feasible, and that this had a positive impact on nutritional status of children of Tamil Nadu. More specifically, these are the findings of the study:

- Program participation: Among children participating in GMP, all service delivery indicators (age at enrolment, regular attendance of sessions, administration of vitamin A, and de-worming), show a substantial increase between 1982 and 1986, though subsequently declined to around their initial levels. Levels of service delivery, however, are generally high.

- Nutritional status: mean weight and malnutrition rates of children aged between 6 and 36 months and participating in GMP have improved over time. Data on non-project areas in Tamil Nadu, and all-India data, that show a smaller improvement over the same time period. Regression analysis of nutritional status on a set of explanatory variables, including the participation in a cotemporaneous nutrition project (the National Meal Program) shows that the latter had no additional benefit on nutritional outcomes. Positive associations are also found between nutritional status and intensive participation in the program, and complete immunization.

- Targeting: using tabulations and regression analysis, it is shown that initially girls have benefited more from the program, but that at the end of the program boys have benefited more. Children from scheduled caste are shown to have benefited more than other groups. Nutritional status was observed to be improving at all income levels, the highest income category benefiting slightly more than the lowest.

INFRASTRUCTURE

Case Study 5: Morocco

Socioeconomic Influence of Rural Roads

Report 15808-MOR, 1996

The Project: The Fourth Highway Project was a US$85 million loan operational from 1983-1990. The project was to improve the national highway system and construct a

freeway from Rabat to Casablanca. It was agreed to also include a major investment in secondary and tertiary roads to help reduce income disparities between regions. The impact evaluation is concerned with the rural roads component of the project.

Methodology: Comparison of four project roads which were all upgraded to asphalt surface and four control roads (roads with no improvement during project period and geographically close to their match). Farm-level surveys were conducted in 9 project villages and 3 in control areas, collecting data from 199 farm households. Village surveys were also applied and qualitative data collected in each community. Since there was no baseline, a ten year recall was used for selected variables. Other data sources, such as the agricultural survey were also used. The report utilizes both single difference (before versus after for control, and project versus control at endline) and double difference estimates. The limitations of the small number of sampled sites and possible differences in initial characteristics are acknowledged in the report.

Findings: A positive impact on a broad range of indicators is identified:

- Traffic-related: (1) after improvement the project roads did not need to be closed during bad weather, whereas they had previously closed for 2-3 months each year imposing greater travel time or restricting travel altogether; (2) traffic volume grew, at a substantially greater rate than the national average for 3 of the 4 roads, with a shift in traffic composition toward heavier freight vehicles; (3) vehicle ownership rose considerably faster in project areas compared to the control and the frequency of public transport increased; and (4) vehicle operating costs fell, reducing commercial trucking rates. Most of these impacts are directly attributable to the project since there is little room for confounding factors. The cost of transport savings is used to calculate the economic return to each of the four roads; these returns are in the range 16-14 percent.

- Agriculture: agricultural activity increased in project areas, with more inputs being used and diversification into fruit trees. New crops, such as sugar beet, were also introduced. By contrast, there was little agricultural innovation in the control areas during the preceding decade. Employment has increased both as large farms expanded operations and because of improved access to urban areas.

- Social services: primary school enrolments grew by 68 percent in the project areas, compared to 61 percent in the control. It became easier to recruit teachers to rural schools with improved roads and absenteeism was reduced. The frequency of visits to health facilities improved by more in the project areas than it did in the control. For example, the frequency of visits to an infirmary increased by 2.7 days per year (from 4.3 to 6) for project areas, compared to 1.1 (35. to 4.6) for the control.

URBAN DEVELOPMENT

Case Study 6: Brazil

Learning from Best Practice in Five Urban Projects

Report No. 16736, 1997

The Project: the study evaluates a package of five urban development projects that operated at different stages in eleven Brazilian cities from 1975-1996 (Medium-size Cities, US$ 70 million; Recife Metropolitan, US$ 108 million; Fortaleza/Salvador Pilot Metropolitan, US$6 million; Parana Market Towns, US$52 million; and North-East Flood Reconstruction, US$ 99 million). In spite of the diversity of project design, these projects shared some common characteristics. (a) They aimed at improving urban infrastructure and services – either upgrading or rehabilitating damaged structures. (b) They included an institutional development component, directed to improve the operation of municipal administrations and planning agencies. (c) They focused on the poor, by targeting poor areas within cities (but not slums), and by implementing small income-generating projects.

Methodology: the study measures the impact of the five projects on three main aspects: (1) living conditions of the poor; (2) decentralization of service provision; and (3) citizens participation in service provision. Three sources of information are used. First, the study compares changes in access to services in study areas to changes in services in other urban areas of the same States, using census data of 1981 and 1991. Second, the study conducted group interviews with beneficiaries who rated a series of indicators before and after the project, and in comparison to other areas of the same city where the project was not implemented. Third, interviews with key informants and on-site visits by experts were used to corroborate and expand the findings obtained from the other data sources.

Findings: (1) The projects improved citizens' access to services and reduced flooding risk. The projects also improved housing conditions. The impact on local economies was found to be modest, while no impact was observed with respect to water supply, sanitation and general maintenance of infrastructure. (2) The projects improved the ability of Municipalities in project preparation, evaluation, and procurement. The project also helped Municipalities to develop a strong ownership of projects that were originally the product of federal government agencies. (3) The project strengthened existing Community Based Organizations (CBO), whose meetings became more regular and more frequently attended after the project. This effect was limited to medium-size cities however, and there was no sign of CBOs empowerment in small cities and metropolitan areas. Women were found to be more active and knowledgeable about urban improvements. The project also encouraged residents to hold municipalities accountable for the services provided.

Case Study 7: Kenya

Development of Housing, Water Supply and Sanitation in Nairobi

Report No. 15586, 1996

The projects: Over a 21 year period (1970-91) the Bank supported five projects, totaling US$120 million, to support water supply and urban development in Nairobi. The three water supply projects aimed to augment sources of potable water supply and strengthen the Water and Sanitation Department. Two other projects upgraded housing units, and other facilities (e.g. health centers), for the poor and expanded sewerage coverage.

Methodology: A survey of 500 households was undertaken in five urban sites, three which had benefited from project interventions and two which had not. Questions were included on current conditions, beneficiary perceptions of project impact, and changes before and after the project. Single and double difference estimates are reported, the latter based on recall. Some participatory analysis was also undertaken.

Findings:

- Water supply has kept with population, but the poor have to pay much more for water obtained through kiosks than do households with direct connections. Sanitation coverage increased, although largely not on account of the Bank projects.
- The supply of housing and affordable rental rooms increased as a result of the project, but home ownership amongst the poor was not affected by the project.
- The projects constructed 11 primary schools and four health centers. Access to health services and primary schools increased, with positive spillover effects from project health and education facilities to non-project areas.
- Problems of flooding and stagnant water increased in all areas, but by much more in non-project areas than in those covered by the project.

Case Study 8: Indonesia

Enhancing the Quality of Life in Urban Indonesia: the legacy of Kampung Improvement Program

Report No. 14747-IND, 1995

The Projects: The study focuses on three components (Kampung Improvement Program, sites and services, and city-wide improvement program) of four urban development projects (Urban I-IV) implemented between 1974 and 1988. Particular attention is paid

to the Kampung Improvement Program (KIP),[33] which accounted for the largest share of funds of the four projects. KIP aimed to provide basic minimum service standards, through roads, footpaths, drainage, garbage collected, water and sanitation, health clinics and primary schools.

Methodology: A household survey was conducted in 9 kampungs: 5 KIP, 2 sites and services and 2 non-KIP. Baseline data were available from other sources, and the coverage of these data was used to guide site selection for the survey. A mixture of single and double difference estimates are given. Qualitative data were also collected. A problem for the study was not being able to collect specific data on the activities carried out in each community.

Findings: The main findings are: (1) improved living conditions in many respects in project areas, but (2) the gap between KIP and non-KIP has closed as non-KIP areas have caught up, and (3) more general improvements in socio-economic status are attributable to good overall economic performance rather than the project. Respondents were asked to identify the sources of changes in living standards. In KIP areas a majority named KIP, whereas in non-KIP the majority attributed changes to the community's own efforts (the other sources were government programs).

More specific findings are as follows:

- Housing quality (walls and floors) was higher in KIP than non-KIP immediately following the project, but subsequently narrowed (closed in the case of walls).
- Major differences between KIP and non-KIP areas are the larger lot size and lower population density of KIP areas.
- Residents of KIP areas are less reliant on water purchased from vendors (an unsafe source), since the majority have access to water piped directly to homes.
- The majority (61 percent) of KIP residents indicated they had no problem with flooding compared to 32 percent in non-KIP areas.
- An uneven picture emerges with respect to education and health outcomes, with no clear beneficial project effect. This finding is attributed to the range of other factors affecting these outcomes.
- Similarly, changing economic fortunes are attributed to factors exogenous to the project.

[33] A kampung is a low-income dense urban area.

Case Study 9: Paraguay

Community-based Rural Water Systems and the Development of Village Committees

Report No. 17923, 1998

The projects: The report covers three rural water supply projects (RWS I – III, a fourth became effective in 1998) amounting to US$ 80 million. The loans financed the work of the National Environmental Sanitation Service (SENASA) with community-level water management committees (*juntas de saneamiento*), which were responsible for operation and maintenance. SENASA created 424 *juntas* in the period under review, of which 257 had become fully-operational covering 400,000 people.

Methodology: Comparisons of health status were made over a ten year period between five villages receiving potable water through the project and five which had not. Data on water use and health were collected from 20 villages covered by RWS I and II, and 130 villages to be covered by RWS IV. These two sources were used for single difference estimates. The evaluation also used beneficiary assessment and a broad range of key informant interviews to apply a theory-based approach.

Findings: During the 20 year period access to safe water under community-based schemes rose from 1 to 20 percent. SENASA data show that 210,000 sanitary units were constructed, though not all of these were under the Bank projects. The *juntas* have been successful in managing water supply, with all of those established still functioning, though a small number were in arrears with SENASA.

Health benefits from the project included the following:

- Hospital visits from villages without potable water were seven times higher than those from villages that had received potable water.
- The study of 150 villages found sickness to be far more common in households without water: the incidence of diarrhea in the last three days was 3.6 times greater and vomiting in the past 15 days 4.4 times greater.
- Unlike in other countries in the region, rural Paraguay did not suffer from a cholera epidemic, which may be partly attributed to better water and sanitation (other factors include health and information).

There has also been a beneficial impact on poverty and income distribution. Those without access to potable water have to pay about four times as much for water which is unsafe. However, this cost savings exists as the *juntas* do not set tariffs at a rate allowing cost recovery, encouraging over-consumption and threatening sustainability. However, even with the subsidy from the project, about 10-15 percent of the population in beneficiary communities had not been able to afford a connection to the system.

Case Study 10: Brazil and the Philippines

Building Institutions and Financing Local Development

Report No. 18727, 1998

The projects: Decentralization initiatives begun in the 1990s put increased responsibilities on local government, though many lack to capacity to effectively carry out their new responsibilities. This study considers two projects with a similar design in two countries: Brazil and the Philippines. Both projects used two main instruments: fiscal and financial reform and investment projects. To obtain a loan for the latter the municipal government had to submit a financial action plan with a comprehensive reform package. There were, however, differences in approach. In Brazil the statewide approach encouraged as many municipalities as possible to participate with a technically simple project. The approach in the Philippines was more selective, allowing a smaller number of municipalities to develop income-generating projects such as public markets.

Methodology: The analysis of municipal finances drew on the large body of information available by virtue of the rigorous reporting requirements for local government. Hence financial data were available for all municipalities. Project and comparison groups were identified on as municipalities participating and not participating in the project respectively in the two states of Brazil and two provinces of the Philippines. These data were used for double difference mean comparisons. In the Philippines a survey was also undertaken of shop and stall holders in one project and one comparison area, the data from which were also used to produce double difference estimates. A survey of mayors was undertaken in one state in Brazil to analyze institutional development aspects.

Findings:

- Participating municipalities performed better than non-participants on financial autonomy; they mobilized more of their own revenue, with property tax responding particularly well to the project. In consequence, participating municipalities did better at balancing their budgets.

- Local economic development, measured by change in trader sales and income and perceived improvements in quality of infrastructure, benefited from the market-place financed by the project (the one sub-project which was subject to impact evaluation).

- The institutional aspects of the project in Brazil were mostly positively perceived, to the extent that beneficiaries promoted the project in other municipalities.

REFERENCES

Baker, Judy (2000) *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners* Washington D.C.: World Bank.

Bamberger, Michael (2006) *Conducting Quality Impact Evaluations under Budget, Time and Data Constraints.* IEG: World Bank, Washington D.C.

Glewwe, Paul William and Margaret Grosh (eds.) (2000) *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of Living Standards Measurement Study* (2 Volumes) World Bank: Washington D.C.

Gupta Kapoor, A. (2002) *Review of the Impact Evaluation Methodologies Used by the Operations Evaluation Department Over the Past 25 Years,* OED Working Paper, Washington, D.C.: World Bank.

Haider S.J, D. Hussain, I. Nayer, A. Maleque, Ashaduzzaman, and A.R. Khan (2004) *Impact Evaluation of Bangladesh Integrated Nutrition Project,* IMED and Read, Dhaka.

Kremer, Michael and Esther Duflo (2005) "Use of Randomization in the Evaluation of Development Effectiveness" in Pitman *et al.*

Levinson, F.J. and J.E. Rohde (2005) Letter to the Editor, *Health Policy and Planning* 20: 405 – 406.

Pitman, George Keith, Osvaldo Feinstein and Greg Ingram (eds.) (2005) *Evaluating Development Effectiveness* New York: Transaction Publishers.

Rawlings, Laura (2005) "Comment" in Pitman *et al.*

Ravallion, Martin (1999) "The Mystery of the Vanishing Benefits: Ms. Speedy Analyst's Introduction to Evaluation" *Working Paper 2153.* Washington D.C.: World Bank.

Ravallion, Martin (2005) "Comment" in Pitman *et al.*

Sack, D.A., S.K. Roy, T. Ahmed and G. Fuchs (2005) Letter to the Editor, *Health Policy and Planning* 20: 406-407. Kenya.

Shadish, William, Thomas Cook and Donald Campbell (2006) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Academic Internet Publishers.

THE WORLD BANK

1818 H Street, N.W.
Washington, D.C. 20433, U.S.A.
Telephone: 202-477-1234
Facsimile: 202-477-6391
Telex: MCI 64145 WORLDBANK
MCI 248423 WORLDBANK
Internet: www.worldbank.org

Independent Evaluation Group
Sector, Thematic &Global Evaluation (IEGSG)
E-mail: eline@worldbank.org
Telephone: 202-458-4497
Facsimilie: 202-522-3125