Introduction: Impact Evaluation in Official Development Agencies

Howard White and Michael Bamberger

1 Introduction

Aid effectiveness has long been disputed. For many years this debate has been fought out at the macro level, though with little consensus. Yet there is also a very large body of evidence from micro studies carried out at the field level of aid-supported projects. What do they tell us about aid effectiveness?

Although, as outlined in Section 2 of this article, project evaluations have been criticised for several biases, a new generation of studies is emerging in official development agencies which are very arguably free of these biases. This issue of the *IDS Bulletin* presents examples of these studies from a number of agencies: Agence Française de Développement (AFD), the Asian Development Bank (ADB), the Inter-American Development Bank (IDB), the Japanese Bank for International Cooperation (JBIC), the Netherlands Ministry for Foreign Affairs, USAID and the World Bank.

Section 3 of this introductory article outlines the methodological challenges to conducting quality quantitative impact evaluations, and Section 4 some of the practical issues involved. Section 5 offers conclusions.

2 Critiques of project evaluations

2.1 Positive bias

There is a widespread view that project evaluations put a positive spin on their findings. Mosley (1987) identified a macro-micro paradox between macro studies which he claimed found no impact from aid on growth and micro studies which largely found a positive impact – for example, that over 80 per cent of all World Bank projects are rated satisfactory. One reason he gave for this paradox was that the micro studies were unreliable as they were biased toward giving favourable findings. More recently Corbridge *et al.* have written that official agencies 'cast their evaluation findings in the best possible light and tread softly around points of contention or criticism' (2005: 3). The sources of this bias are usually identified to be either direct censoring by agency staff or self-censoring by consultants employed as evaluators, as they don't wish to jeopardise future contracts by being unduly critical; Teller's article in this issue suggests such problems can exist in USAID.

Another source of bias occurs in cases where information is only obtained on project participants and there is no control group – this is the case for the very large number of 'impact evaluations' in which budget, time or sometimes political influences limit the use of comparison group data. Due to the project selection bias discussed later, project participants are more likely to be successful with respect to whatever the project is trying to achieve (e.g. school enrolment, development of small businesses, improved nutritional behaviour) than non-project groups, so interviewing participants only is likely to over-estimate project outcomes. Selection bias also arises from attrition between the rounds of a panel survey (Ito, this *IDS Bulletin*). There is also a sample selection bias in which projects are chosen to be evaluated as agencies will not devote evaluation resources to projects which are known to have failed.

There is something in this last argument. There is no point in launching a full scale impact evaluation if it's known that the project didn't even get very far in producing the desired outputs. A recent, well designed evaluation system for a nutrition project in Bangladesh was abandoned for precisely these reasons, it being decided instead to use action research to analyse the reasons for failure. If the agency has a comprehensive system of project reviews, then such failures will show up in these

numbers. Unfortunately, such comprehensive systems are not yet the norm (see White 2005).

But the other accusations of bias from desiring a positive spin are frequently touted but rarely backed up. Drawing on our own experience we can say that such pressures are rare exceptions; indeed some agencies actively encourage critical findings. This is of course anecdotal evidence, but so is that of the critics – there has not been to our knowledge a systematic analysis of this issue. However, one only need look at evaluations themselves to see that these can often be critical. In IEG's recent set of impact evaluations we have pointed to the low rate of return to irrigation investments on account of construction delays and cost overruns, the inferior return to off-grid electrification compared to grid electrification and so questioned the emphasis on the former, and heavily criticised a nutrition programme for not being cost effective. It is true that operational staff in the World Bank opposed these findings, but IEG published the results unaltered. This is not to say that such pressures do not exist. A useful distinction is that between self-evaluations (commissioned by the project or agency operational plan) and those undertaken by evaluation departments which have varying degrees of independence.

2.2 Short-term bias

Another criticism has been that project-supported interventions are not sustainable, so that evaluations made during or immediately after the project are misleading. This is indeed a good point. The value of an investment is much less if the benefit stream is not sustained into the future. But it is not a point on which impact evaluation can historically be faulted, as in many agencies impact studies are by definition studies carried out some years after the intervention has closed. This point of view is still reflected in the DAC definition of impact as 'Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended'. JBIC's study of the Jamuna Bridge in Bangladesh, presented in the article by Ito *et al.* in this *IDS Bulletin*, falls into this tradition, being carried out five years after completion of the bridge (there was a baseline shortly before the bridge opened).

But for the reasons explained below, the meaning of impact evaluation has shifted and more studies are done during the project or shortly after completion. However, a good impact evaluation design will address this question of sustainability by ensuring that the necessary conditions are in place. One of the benefits of using a programme theory (discussed later in this article) as part of the evaluation design is that a well articulated theory model can help define the time horizon over which outcomes and impacts are expected to be achieved. This can caution against the use of unrealistically short time periods for the impact evaluation, and can provide a useful tool for evaluators who need to convince clients why the time horizon for the evaluation cannot be shortened without running the danger of producing the misleading conclusion that the interventions being assessed 'do not work.'

2.3 Sample coverage bias

Many impact evaluations (as well as other types of evaluation) use an easily-available administrative list or map as the sampling frame for selecting project and comparison group samples. In many cases these lists are not complete as when, for example, they only include families who have registered to receive benefits (for example the internally displaced, the poorest or female-headed households), or who have land or property titles. Often the groups who are not included are the poorest, most vulnerable or people with at best an ambiguous legal status. Consequently the samples selected for the evaluation will have a positive bias as the groups most in need or least likely to succeed have been omitted.

2.4 Beneficiary selection bias

The DAC definition of impact given above implies a counterfactual analysis: what happened to outcomes with the project, compared to what they would have been in the absence of the project. The 'without' case has usually been captured by a control group, which should more properly be called a comparator group as the evaluator cannot control what happens to this group. The comparator group should be identical to the beneficiaries (the treatment group) in all respects except that they don't receive the intervention. This has usually been done by taking neighbouring areas or matching treatment and comparator communities on available socio-economic data. However, the way in which beneficiaries are selected may undermine the comparability between treatment and comparison groups.

International development projects typically use one of two procedures for participant selection: self-

selection (people are invited to apply, for example, for small business loans, or communities apply to participate in a programme to provide water, schools or other social services – as in the case of the social funds discussed in the articles of Ruprah, and Ito *et al.* in this *IDS Bulletin*) and administrative selection (the project implementing agency selects the individuals, communities or administrative areas who will participate). Hence participants are likely to have special characteristics, often correlated with project success, which distinguish them from non-participants. In econometric terms, this is a problem of endogeneity which will bias the impact estimates. This bias is illustrated in AFD's impact study of microfinance in Madagascar, where an earlier impact study had found a positive impact but this was not supported by a second, more rigorous, impact design (Naudet and Delarue, this *IDS Bulletin*). By contrast, a 'naïve comparison' for an IDB-supported Social Investment Fund found poverty to have risen amongst beneficiaries, whereas a rigorous impact evaluation design showed a significant impact on poverty reduction (Ruprah, this *IDS Bulletin*).

If selection characteristics are known and observed then they can be controlled to remove the bias by using a range of quasi-experimental (regression-based) techniques. But if selection characteristics cannot be observed – depending on such things as 'entrepreneurial' or 'community' spirit – then the omission of these variables will bias regression-based estimates of project impact. However, in the cases that these unobserved determinants do not vary over time (time invariant) then their influence can be removed by double differencing (the difference in the change in the outcome for the treatment and control groups), and so selection bias is eliminated – but we have to assume this time invariance as it can of course not be observed. But if the time invariance of unobservables cannot be observed then the preferred approach would be an experimental design, also known as a randomised control trial (RCT).

Failure to address this problem has been a growing criticism that there has been very little proper impact evaluation, notably from the Poverty Action Lab at Harvard and the report of the Center for Global Development entitled 'When Will we Even Learn?' (CGD 2006). It is true that many older impact evaluations had control groups which did not explicitly allow for sample selection bias, so may (but also may not) have obtained biased estimates. But, other than randomisation, the techniques which have become most popular for dealing with these problems (propensity score matching and regression models adjusting for sample selection bias) have only been developed in the last 20 years. In response to these criticisms, impact studies are now very concerned to establish a 'credible counterfactual', meaning one which addresses the selection issue (and other possible sources of bias, see White 2006). The next section addresses how to go about this.

To do this, we draw on the experiences of members of the Network of Networks on Impact Evaluation (NONIE). NONIE, created in November 2006, brings together the members of evaluation networks of the official development agencies – the DAC Evaluation Network, the UN Evaluation Group (UNEG), and the Evaluation Cooperation Group (ECG) of the multilateral development banks. NONIE is the framework under which official agencies are collaborating to improve the quantity of quality impact evaluations they produce. A separate initiative, the International Institute for Impact Evaluation (3IE), has grown out of the work of CGD, which will be an independent agency commissioning impact studies on enduring questions of interest to the development community – 3IE is not yet operational at the time of writing (November 2007), but is expected to be so by mid-2008.

3 Designing an impact evaluation

Approaching an impact evaluation, the first design decision is whether the intervention is amenable to a quantitative approach. When making this decision it is important to remember that a large part of the evaluation community and many development agencies (particularly but not exclusively NGOs) would question whether a quantitative approach is the best, or even an appropriate approach, for understanding the effects of development programmes operating in complex and culturally diverse settings.[1] Evaluators, including those from a quantitative background, would do well to understand the arguments for and against both quantitative and qualitative approaches before starting to assess whether conditions permit the use of a quantitative evaluation design.

Assuming the decision is made to try to use a quantitative design, the general guiding criterion is the number of observations (n) which will be available; a small n means econometric techniques

will not be applicable. For a small n (less than 30 or so) sample, then a case study approach is likely to be more appropriate, relying on different means of establishing links between inputs and impacts. There is also an argument that quantitative techniques cannot be used when the outcomes are non-quantifiable, but there is a growing body of work measuring such apparent 'immeasurables' as empowerment.

For a quantitative design, there is considerable advantage in designing an impact evaluation at the start of the intervention and collecting baseline data; this is called an *ex ante* design. A randomised approach is simply not possible unless the evaluation is put in place at the outset. But having baseline data also allows the construction of a panel design that permits the application of a 'difference in difference' or 'double difference' approach; that is, comparing the change in outcomes in the treatment area compared to the comparator group. As explained above, this approach can control for time invariant unobservables. Impact evaluation designs are therefore stronger if put in place *ex ante*. As the experience of the Agence Française de Développement (AFD) described in the article by Naudet and Delarue in this *IDS Bulletin*, shows, having an evaluation department committed to rigorous impact evaluation makes this more likely to happen, an argument explicitly stated in the title to the IDB article: 'you can get it if you want it' (Ruprah, this *IDS Bulletin*). In the case of the World Bank, the research department created the Development Impact Initiative (DIME) which provided project managers with support (workshops, some seed finance and links to researchers for design) which has led to a blooming of *ex ante* evaluation designs (as of October 2007, 160 were underway and a further 60 planned). By contrast, as documented by Teller in this *IDS Bulletin*, changing management fashions in USAID have compromised evaluation quality.

However, evaluators are often faced with designing an impact study *ex post*. Even then, it may be that there are existing data available which can serve as a baseline, so a double difference may be applied. But if that is not possible, then the design has to rely on a quasi-experimental approach, which denotes a range of means for constructing a comparator group. The fact that there is a comparator group means that data collection needs extend beyond the beneficiaries of the intervention. Although there are cases in which causation is so obvious that no control is necessary, this is the rare exception rather than the norm. Quasi-experimental methods are not described in detail here, though examples are to be found in the articles in this *IDS Bulletin*; for further discussion of approaches see IEG (2006), Baker (2000) and Ravallion (2001). Suffice it to note that quasi-experimental approaches are regression-based, so that data collection need cover not only participation and outcomes but also determinants of both participation and outcomes.

We present two figures to assist in evaluation design choices. First is a decision tree, guiding the evaluator as to how to choose the best design given the constraints they face. The decision tree is related to interventions for which a quantitative approach is both feasible and appropriate. The second figure outlines the main features of different design options, ranked by quantitative robustness.

3.1 Ensuring policy relevance

The impact evaluation design should also ensure policy relevance: it should be able to answer not just what works but also why (or why not, as the case may be). Relevance is best assessed by using a theory-based approach, which maps out the causal chain from inputs to outcomes/impacts. The advantage of a theory-based approach is that it allows identification of problem areas in programme design and implementation which have hindered the achievement of outcomes. The articles by Ruprah, and by Naudet and Delarue in this *IDS Bulletin* argue against simply measuring impact without seeking to understand why. Examples also come from IEG's work. For example, a study of agricultural extension in Kenya found that extension advice promoted techniques which had already been widely adopted, so the finding of little impact on yields was hardly a surprise (World Bank 1999). Evaluation of a nutrition programme in Bangladesh found a number of weak and missing links in the causal chain, which both explained the project's low impact and pointed to needed changes in project design. Such findings clearly have greater policy usefulness than simply finding that a programme has little or no impact.

The theory-based approach 'opens the black box' to allow observation of how the project is actually implemented on the ground as compared to what was planned in the operations handbook. A pure impact estimate does not help us to understand how

Figure 1 **Decision tree for selecting evaluation design to deal with selection bias**

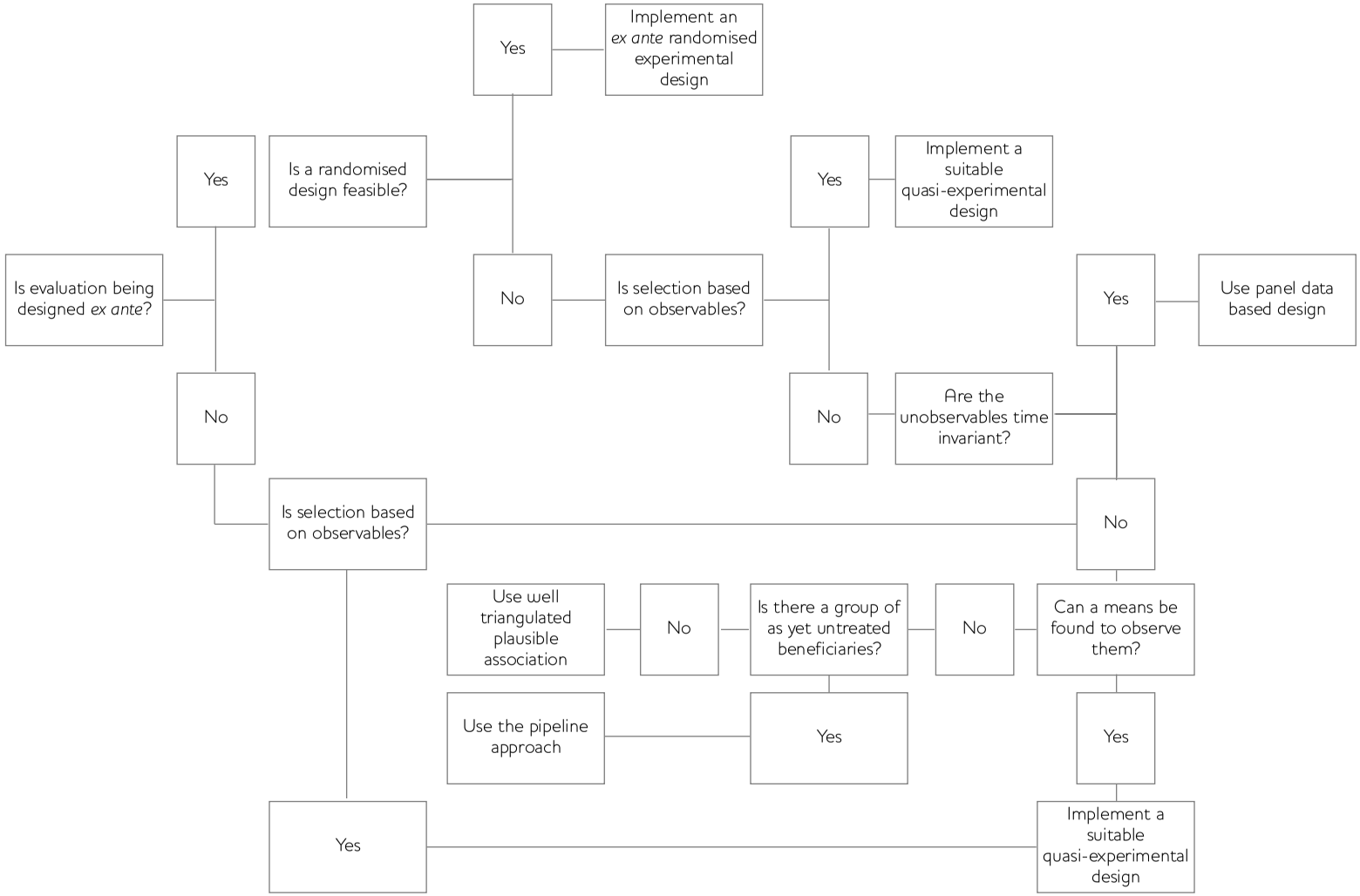

Figure 2 **Eight commonly used quasi-experimental and non-experimental impact evaluation designs**

Key T = Time P = Project participants; C = Control group P_1, P_2, C_1, C_2 First and second observations X = Project intervention (a process rather than a discrete event)	**Start of project [pre-test]**	**Project intervention [Process not discrete event]**	**Mid-term evaluation**	**End of project [Post-test]**	**The stage of the project cycle at which each evaluation design can to be used**
Quantitative impact evaluation design	T_1		T_2	T_3	
Relatively robust quasi-experimental designs					
1. ***Pre-test post-test non-equivalent control group design with statistical matching of the two groups.*** Participants are either self-selected or are selected by the project implementing agency. Statistical techniques (such as propensity score matching), drawing on high quality secondary data used to match the two groups on a number of relevant variables.	P_1 C_1	X		P_2 C^2	Start
2. ***Pre-test post-test non-equivalent control group design with judgmental matching of the two groups.*** Participants are either self-selected or are selected by the project implementing agency. Control areas usually selected judgmentally and subjects are randomly selected from within these areas.	P_1 C_1	X		P_2 C_2	Start
Less robust quasi-experimental designs					
3. ***Pre-test/post-test comparison where the baseline study is not conducted until the project has been underway for some time*** (most commonly this is around the mid-term review).		X	P_1 C_1	P_2 C_2	During project implementation (often at mid-term)
4. ***Pipeline control group design.*** When a project is implemented in phases, subjects in Phase 2 (i.e who will not receive benefits until some later point in time) can be used as the control group for Phase 1 subjects.	P_1 C_1	X		P_2 C_2	Start
5. ***Pre-test post-test comparison of project group combined with post-test comparison of project and control group***	P_1	X		P_2 C_2	Start
6. ***Post-test comparison of project and control groups***		X		P_1 C_1	End
Non-experimental designs (the least robust)					
7. ***Pre-test post-test comparison of project group***	P_1	X		P_2	Start
8. ***Post-test analysis of project group***		X		P_1	End

Source Adapted from Bamberger, Rugh and Mabry (2006).

well the intervention is working (or not working) in different contexts. This is often expressed by saying that when an intervention does not achieve its intended outcomes, a pure impact estimate cannot distinguish between 'design failure' and 'implementation failure'. A similar point is made by proponents of 'realistic evaluation', who point to the fact that the findings from impact studies of the same intervention have often given mixed findings, which is not helpful to policymakers (e.g. Pawson and Tilley 1997). However, they argue that these mixed findings are hardly surprising given the different contexts in which interventions take place. There are some clear implications from this argument: (1) the evaluation design needs to be aware of context and, to the extent possible, collect data on it; (2) where data are available, the impact estimation approach should allow for context; and (3) attempts to generalise about impact need to be context-specific (and consequently less ambitious than is often the case).

Addressing selection bias provides a stepping stone toward a theory-based approach, since the selection equation allows for an analysis of targeting and possible constraints on participation. IDB finance to publicly supported mortgage programmes has been criticised for high delinquency rates, critics arguing that the programmes should be transferred to the private sector which doesn't suffer from such delinquency. But the selection analysis shows that the clients are very different, with the clients of public programmes more likely to face difficulties in meeting repayments – a problem addressed by increasing the grant component rather than switching provider (Ruprah, this *IDS Bulletin*).

Relevance can also be enhanced by the use of mixed methods, which isthe issue taken up in White's article in this collection. He provides both examples in which qualitative material, including fieldwork by the study team, pointed to further quantitative analysis which resulted in clearer and more focused policy conclusions, and examples in which quantitative findings have led to qualitative work which has informed the policy response. Examples of the former include the important role played by the mother-in-law household decision-making in Bangladesh, leading to the recommendation that they also be targeted by nutritional counseling for child health; and the growing dichotomy in Ghana between schools in poor and better off areas, hence the proposal for a central fund for school development targeted at poorer schools. As an example of the latter, a second round survey of a project supporting women's self-help groups with credit and training found significant drop-out from the first to the second round. Qualitative methods, such as oral life-histories, were used to explain the reasons for these drop-outs, and helped to recommend the appropriate policy response.

De Kemp's article in this *IDS Bulletin* illustrates the lessons which can be learned from a well-contextualised approach. Both Uganda and Zambia, especially the former, have rapidly expanding primary school enrolments. Initially there was dramatic overcrowding. But subsequent teacher recruitment has brought down teacher-pupil ratios, though they remain unacceptably high. Nonetheless, critics of universal primary education argue that quality suffers. However, de Kemp's analysis shows that average test scores amongst those groups already in school before the expansion have *not* suffered. Overall average performance has, unsurprisingly, been dragged down by the enrolment of children from less privileged backgrounds less accustomed to schooling. The study points to a number of ways in which standards can be raised: reducing teacher absenteeism as part of a general strategy to increase contact time, teacher training to improve textbook usage, and improved school management.

Finally, the requirements of the evaluation may require changes in project design. This is most obvious in the case of RCTs which have clear implications for project implementation arrangements. In AFD's evaluation of health insurance project design changes had to be during implementation to increase the chance of a project impact – specifically carrying out more information campaigns at village-level and delaying introduction in other villages (this was a RCT but the same factors would have applied to a quasi-experimental design; Naudet and Delarue, this *IDS Bulletin*).

4 Practical issues in evaluation design

Evaluators are often faced with both time and budget constraints, so the challenge is to produce a rigorous design whilst working within these constraints.

4.1 Addressing budget constraints

Impact evaluations are seen as being more expensive than other types of evaluation, frequently, though not always, with good cause. In AFD, just two

impact evaluations are accounting for 25 per cent of the unit's total budget (Naudet and Delarue, this *IDS Bulletin*), and in IEG each impact study has cost around US$350,000.[2] By contrast, the Inter-American Development Bank has been very successful in supporting low-cost impact evaluations averaging less than US$50,000 each while avoiding any new data collection (Ruprah, this *IDS Bulletin*). It has used proposals submitted to undertake studies to identify existing data sources, to which it can obtain access for local research teams who may not otherwise be able to obtain those data for analysis (and in consequence, who are able to put in cheap bids to achieve this privilege).

Five options can be considered to reduce the budget (Bamberger *et al.* 2006: Chapter 3). First, considerable cost savings are often possible by eliminating one or more of the four data collection points (pre-test/post-test project and control group). For example, design 5 eliminates baseline control group data and design 6 eliminates all baseline data. There is clearly a trade-off that must be assessed for this and the following options between cost savings and methodological rigour. Second, the data collection instruments can be simplified to reduce the amount of information to be collected. In other cases it may be possible to reduce the number of people from whom information is collected. Third, the creative use of secondary data can often reduce data collection costs. Fourth, a judicious assessment of expected effect size and power analysis may sometimes make it possible to reduce sample size while still obtaining satisfactory estimates of project impact. Finally, there are often ways to reduce the costs of data collection. One possibility is to use less expensive interviewers such as medical students or student teachers rather than commercial interviewers. But a point comes at which, if the budget is too tight, it is best to abandon plans for a rigorous quantitative impact evaluation and ask instead how the money may best be spent to serve the evaluation functions of learning and accountability.

While it is often assumed that the evaluation will always require the collection of primary data, it is often possible to significantly reduce time and cost, as well as enhance quality by drawing on available secondary sources of data (IEG 2006). In addition to primary data collection in both project and control areas, it may be possible to obtain data from an existing or planned survey. For example, an evaluation can 'piggyback' on a planned survey, paying for an additional module and, if necessary, oversampling of the project area.

At the level of the agency there are two alternative approaches which address the budget issue. One is to shift responsibility for impact evaluation to the research department. This debate on location took place in AFD, though evaluation and research fall under the same department making it less of an issue. Some years ago IEG decided to cease IEs on the grounds of expense, and leave them for the research department, but then initiated the current programme at the request of the Board. We would note that research departments do not have the same mandate as evaluation departments so that the impact studies they produce may not answer the questions of most interest to policymakers.

The second possibility is to integrate impact evaluation into larger studies. This approach is currently being considered in AFD, and has been tried with varying degrees of success in recent years in IEG. IEG's experience shows that most staff do not have the required skills, so the in-house IE expertise needs to devote considerable time to the activity, even if it is just to manage the IE component being carried out by consultants. The analysis of Forss and Bandstein in this *IDS Bulletin* points to the possible limitations of this approach – they argue that the terms of reference for evaluation studies contain multiple points, few of which are covered by a traditional rigorous impact evaluation design. In consequence, the design is dominated by the need to answer the many other questions raised by the donors, any focus on a rigorous approach to attribution falls by the wayside and so in the end the answer to the impact question is fudged or simply ignored altogether. Hence they argue that impact studies need to be carried out as a separate product line rather than merged with other evaluations. In response, we certainly support the view that evaluation has many functions other than impact assessment, but would argue that a good theory-based design can answer questions of both process and impact.

4.2 Addressing time constraints

Impact evaluations suffer from two types of time constraint. First is the usual issue of too little time being allowed for the study, which we discuss below. But IEs also suffer since the most rigorous designs

require involvement from the design stage of the project. Hence, as argued by Forss and Bandstein in this *IDS Bulletin*, evaluation needs to be more closely integrated into the project planning stage. The two ongoing AFD impact evaluations needed 18 months preparation prior to the start of the project (Naudet and Delarue, this *IDS Bulletin*).

The skills constraint applies to both the staff of the evaluation department and to the consultants usually employed by that department. This lack of skills may create a barrier to implementing rigorous impact evaluations, as appears to have happened in USAID. In response to lack of in-house capacity the department can either hire new staff with specific skills (as AFD and IEG did) or undertake training. In response to the constraint amongst consultants, the department can of course form relationships with new partners with the requisite skills – as both AFD and IDB have done – or encourage skills development amongst its traditional partners – as both NORAD and BMZ are doing. However, these new partners will be less in tune with the usual evaluation questions and seek to take the work in a more academic direction than desired by the agency. Hence the need to ensure policy relevance, for both the agency but also of course for local policymakers.

Policy relevance is just one ingredient for the desired aim of policy impact. Local stakeholder involvement is another – both at the policy level and amongst those implementing the intervention under study. Involvement of project staff is of course vital, if only to gain access to project sites. But their buy-in is also needed to help preserve the integrity of the evaluation design and should the evaluation demand in-course correction. For example, AFD's micro-finance study determined that greater sensitisation efforts were required to increase coverage to chance of the study uncovering statistically significant impacts, and, for the same reason, the rolling out of the programme to the control areas was delayed by a year.

Most of the above techniques for addressing budget constraints within individual studies can also be used to reduce time (Bamberger *et al.* 2006: Chapter 4). When time is a constraint but there is an adequate budget it is sometimes possible to contract local consultants to conduct preparatory studies. This increases the efficiency of the limited time expensive foreign or out-of-town consultants have available for in-country or project visits. Video-conferencing can also be an effective way to improve coordination and save time. Hiring more researchers, interviewers or data analysts may also be considered to reduce the time required for data collection and analysis. However, increasing the size of the research team also increases the complexity of coordination so less time may be saved than expected. Data collection technology such as hand-held computers, internet surveys and optical scanning are also possible time-savers.

4.3 Addressing data constraints

Real-world evaluations often lack baseline data, particularly on the control group but also quite often on the project population as well. Where selection is based on unobservable factors that don't vary over time, the lack of a baseline is especially important because the influence of these factors can be removed by double differencing if good baseline data is collected. For the same reason, double differencing also helps when there has been inadequate definition of the control population. A number of strategies are available to reconstruct baseline data (Bamberger, Rugh and Mabry 2006: Chapter 5).

First, as mentioned above, an existing survey may serve this purpose. Second, existing documentary data from within the organisation or from other sources can be used, or key informants can also be asked to provide information on pre-project conditions. Finally, informants can be asked to recall their situation prior to the start of the project. Some evaluators question the validity of recall as it is particularly vulnerable to bias because of intentional distortion or lapses of memory. But all questionnaires are based on recall – so it is actually a question of degree rather than whether the approach should be used at all. Areas such as income and expenditure and fertility behaviour, in which extensive research has been conducted on the reliability of recall, have shown that it is possible to identify the direction and magnitude of bias as well as identifying ways to reduce the bias. Major events and purchases (such as main assets like a vehicle or livestock) can be recalled with reasonable accuracy, especially if other methods are used to triangulate the information. Asset measures, combined with indicators of housing quality, are increasingly used as a proxy for the more difficult to measure outcome of household income. Krishna *et al.* (2006) use recall for an asset-based approach to analysing poverty

trends in a number of Indian villages over a 25 year period.

There are also a number of PRA techniques that can be used to reconstruct baseline conditions. The term PRA (Participatory Rural Appraisal) is now commonly used as a generic term to describe a wide range of participatory planning and evaluation techniques that are used with groups or communities to identify their development priorities; their perception of the constraints affecting the achievement of their goals and the resources they can draw on; and their opinions on the effectiveness of community organisations and external programmes. PRA techniques were originally developed, drawing heavily on the work of Robert Chambers (e.g. Chambers 1994a, b and c), for working with mainly rural communities with low levels of literacy and often with difficulties in expressing their ideas verbally and consequently PRA has developed a wide range of techniques that do not involve reading or writing and that use non-verbal communication. With all of these techniques a facilitator works with community groups, rather than individuals, and uses social maps, charts and other visual and easily understandable techniques to reconstruct time-lines, trend analysis, historical transects and seasonal diagrams to trace the evolution of the community and the critical incidents in its history (Kumar 2002). PRA methods are also helpful for addressing other data constraints which occur when data collection methods are not adequate for collecting sensitive information or for identifying, locating and interviewing difficult-to-reach groups. In addition to questions concerning potential biases in information collected from groups and questions as to how the data can be incorporated into quantitative analysis, a problem with most group-based data collection is that the sample size is significantly reduced as the unit of analysis becomes the group rather than the individual or household. This is particularly important when group-based techniques are advocated as a way to reduce the costs of data collection through household sample surveys.

5 Conclusions

Aid effectiveness has long been disputed. Over four decades of analysis at the macroeconomic level have been inconclusive. But there is a growing body of evidence from detailed, microeconomic field-level impact evaluations. As shown by the articles in this *IDS Bulletin*, the design of these studies increasingly addresses the different sources of bias complained about by critics of official agencies' evaluations. These evaluations therefore provide a firm basis for drawing conclusions on aid effectiveness, though a greater number are required to permit generalisation – but generalisations should always bear in mind the specific context under which an intervention has or has not worked.

Challenges remain. The scale of studies to date has been small, though a number of initiatives exist to change this situation. The technical skills required are demanding and are not widely available in official agencies or developing country governments. This situation is also changing, but will require deliberate action. Finally, there is some resistance to widespread adoption of these techniques. What is required is a greater understanding of both the scope and limitations of quantitative impact evaluation. They are not always appropriate. Nor are they the only type of evaluation which should be utilised. But greater use of well designed, theory-based, rigorous impact evaluations will enhance the likelihood of achieving international poverty reduction targets.

Notes

1 For a review of the arguments for and against quantitative and qualitative evaluation designs see Bamberger *et al.* (2006: Chapters 11 and 12).

2 Though this is toward the lower end of the typical budget for an IEG 'large study', these large studies cover whole sectors or programmes rather than a single intervention as impact studies have traditionally done.

References

Baker, J.L. (2000) *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*, Washington DC: World Bank

Bamberger, M.; Rugh, J. and Mabry, L. (2006) *Real World Evaluation: Working Under Budget, Time, Data and Political Constraints*, Thousand Oaks, CA: Sage

Chambers, R. (1994a) 'The Origins and Practice of Participatory Rural Appraisal', *World Development* 22.7: 953–69

Chambers, R. (1994b) 'Participatory Rural Appraisal: Analysis of Experience', *World Development* 22.7: 1253–68

Chambers, R. (1994c) 'Participatory Rural Appraisal: Challenges, Potentials and Paradigm', *World Development* 22.7: 1437–45

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

Corbridge, S.; Williams, G.; Srivastava, M. and Véron, R., (2005) *Seeing the State: Governance and Governmentality in India*, Cambridge: Cambridge University Press

Krishna, A.; Lumonya, D.; Markiewicz, M.; Mugumya, F.; Kafuko, A. and Wegoye, J. (2006) 'Escaping Poverty and Becoming Poor in 36 Villages of Central and Western Uganda', *Journal of Development Studies* 42.2: 346–70

Kumar, S. (2002) *Methods for Community Participation: A Complete Guide for Practitioners*, Rugby: ITDG Publishing

Mosley, P. (1987) *Overseas Aid: Its Defence and Reform*, Brighton: Harvester Wheatsheaf

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*, London: Sage

Ravallion, M. (2001) 'The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation', *World Bank Economic Review* 15.1: 115–40

White, H. (2006) *Impact Evaluation Experience of the Independent Evaluation Group of the World Bank*, Washington DC: World Bank

White, H. (2005) *Challenges in Measuring Development Effectiveness*, IDS Working Paper 242, Brighton: IDS

World Bank (1999) *Agricultural Extension: The Kenya Experience*, Washington DC: World Bank