



ROOM DOCUMENT 5

DAC Network on Development Evaluation

IMPACT EVALUATION: AN OVERVIEW AND SOME ISSUES FOR DISCUSSION

This note has been prepared by the Independent Evaluation Group of the World Bank, in collaboration with the DAC Secretariat, for discussion at the 4th meeting of the DAC Network on Development Evaluation, 30 – 31 March 2006.

**4th meeting
30 – 31 March 2006**

Introduction

The need to show results

The results agenda and the Millennium Development Goals (MDGs) have increased focus on final development outcomes such as income poverty and under-five mortality. Governments and development agencies want to be able to demonstrate that their support is having an impact on these outcomes. Often evidence of such a link between money and outcomes is weak or missing. A recent report states that there is an “Evaluation Gap” (Center for Global Development, 2005), claiming that only a handful of reports produced yearly that can scientifically attribute changes in social outcomes to particular projects or programmes. More impact evaluations are said to be required to fill this gap.

There is also clearly some confusion about the meaning of impact evaluation and this paper aims to clarify some of the concepts which may hopefully contribute to facilitate the dialogue in the current debate on impact evaluations.

What is impact evaluation?

Impact evaluation has taken different meanings during the last twenty years. The DAC Principles of Evaluation of Development Co-operation defined impact as one of the key five evaluation criteria already back in 1991. The five evaluation criteria contained in the 1991 Principles, and confirmed in the 1998 review, are today applied by most development agencies as the standard evaluation criteria. Impact is defined in the Principles as: “The positive and negative changes produced by a development intervention, directly or indirectly, intended or unintended”. This involves the main impacts and effects resulting from the activity on the local social, economic, environmental and other development indicators. The examination should be concerned with both intended and unintended results and must also include the positive and negative impact of external factors, such as changes in terms of trade or financial conditions.

The following four meanings of impact evaluation have recently been the most common:^{1,2}

- An evaluation which looks at the impact of an intervention on final development outcomes, rather than only at project outputs, or a *process evaluation* which focuses on implementation;
- An evaluation concerned with establishing the counterfactual,³ 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 in order to examine its long-term effects; and
- An evaluation considering all interventions within a given sector or geographical area.

1. Whilst preparing the DAC Glossary of Key Terms in Evaluation, the DAC Secretariat collected 15 definitions of impact evaluation from different development agencies. Impacts in the glossary are defined as meaning essentially various types of effects (direct, indirect, intended, unintended etc).

2. A fifth use and perhaps the most common use of the term is in impact assessments; which are increasingly used in a variety of ex-ante appraisals, e.g. poverty impact assessment; conflict assessments, environmental impact assessment etc. While these are ex-ante appraisals, the term illustrates the multi-dimensional nature of the use of the impact concept in development agencies today.

3. “The situation or condition which hypothetically may prevail for individuals, organizations, or groups were there no development intervention” (DAC Evaluation Glossary).

These four definitions are not mutually exclusive. Increased results-orientation has given prominence to the first of these definitions – a focus on development outcomes. But there is increased awareness that attention need be paid to attribution through the establishment of a good counterfactual.⁴

But the different meanings do not necessarily coincide. Final development outcomes need not necessarily be long-term effects. A feeding program for pregnant women to reduce the incidence of low-birth weight must have its impact in less than nine months (though it may be the case that the strength of this impact changes over time). However, other benefits do indeed take longer to be realized – e.g. agricultural diversification from improved market access or irrigation – so that the evaluation time frame needs to allow for this fact.

The need for results supports the idea of seeing impact evaluation as being concerned with final development (or welfare) outcomes. There has been a rise in the sophistication of quantitative techniques available to measure impact in the last decade – often called rigorous impact evaluation.⁵ The need for more impact evaluation is a need for more, more rigorous, impact evaluation. But, as argued below, these need to be well-contextualized, policy-relevant studies adopting methodological flexibility to fit the conditions under which they are carried out.

The place of impact evaluation

But a sense of perspective is needed. Impact evaluation is just one of a range of types of evaluation. Valuable lessons come from other evaluations, and this will continue to be the case. The call is for more impact evaluation, and that should be supported. But this does not mean impact evaluation should be done at the expense of all other types of evaluation.

Approaches to impact evaluation: an overview

Different approaches may be taken to impact evaluation. In brief, these are:

- Quantitative impact evaluation: analysis based on a representative survey of treatment group and a comparison group, preferably both before and after the intervention. There are a range of possible techniques under this heading (experimental and quasi-experimental).
- Participatory impact evaluation: analysis based on participatory methods amongst beneficiaries.
- Theory-based (program logic) approaches: analysis tracing the log frame from inputs to outcomes, using a mix of methods to establish causal linkages.

Although these are different approaches, good impact evaluations will use a mixture. Qualitative analysis, even if not a full-blown participatory analysis, can help provide valuable context. A theory-based approach helps build the story around the intervention and understand why it worked or not. But

4. The World Bank PovertyNet website follows the first of the four definitions – “an impact evaluation assesses the changes in the well-being of individuals that can be attributed to a particular project, program or policy” – but makes it clear that counterfactual analysis is required to qualify as impact evaluation in their terms. OED’s current program of impact evaluation similarly carries out counterfactual analysis of final welfare outcomes. CGD’s ‘Evaluation Gap’ document identifies impact evaluations as studies which “document whether changes in social outcomes can be attributed to particular programs”, i.e. a counterfactual analysis of outcomes.

5. This expression is used by both IEG in its recent note on impact evaluation (OED, 2005) and in the CGD paper.

quantitative methods generally give a more authoritative indication of the counterfactual impact on outcomes.

Approaches to Impact Evaluation

*Quantitative impact evaluation*⁶

The cornerstone of quantitative impact evaluation is data collection from a statistically representative sample using a structured questionnaire. It is strongly preferable that data are collected both before the intervention (baseline) and after (endline). A midterm survey is also an advantage. Data should be collected from both the affected population (the treatment group) and a comparison group.⁷ Project impact is then calculated as either a single difference (difference in outcome between project and control after the intervention), or double difference (the difference in the change in outcomes for the project and control before and after the intervention).

Selection of an appropriate comparison group is one of the main challenges in impact evaluation. They should be identical to the treatment group except that the latter receive the intervention and the former do not. In practice this is difficult to achieve for two reasons. First, beneficiaries of the intervention may be selected (or self-select) on the basis of certain characteristics. If these characteristics are observed then a comparison group with the same characteristics can be selected. But if they are unobserved then in principle only a randomized approach can eliminate selection bias.⁸ Second, the comparison group may be contaminated either by spillover effects from the intervention or a similar intervention being undertaken in the comparison area by another agency.

The main approaches are as follows:

- Experimental (randomized) approaches: Experimental (or randomized) evaluation design tackles the first, but not the second, of these problems. Experimental design requires that the eligible population be identified and then a random sample of those treated. 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. Experimental methods have the strong appeal of avoiding an otherwise unknown bias from selectivity, but are in practice only applicable to a narrow range of the interventions supported by development agencies. Where they are applicable then they should be used, certainly more so than is done at present.
- 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. Clearly the approach can only be used for activities which continue beyond the end of the project being evaluated.

6. A longer version of this sub-section may be found in the annex to this note.

7. Usually referred to as the control group. Strictly speaking this term is only correct in an experimental design in which after factors are controlled so as not to vary between treatment and control.

8. This statement needs to be qualified in two ways. First if the unobservables are time invariant they can be removed by double differencing. Second, if the direction of the bias can be reasonably assumed then upper or lower estimates of project impact can be made.

- 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. PSM uses statistical modeling to identify a group of individuals, households or firms with the same observable characteristics as those participating in the project. The potential problem with PSM is that facing all quasi-experimental approaches: selection on unobservables. Unobservables which simply affect project outcomes and 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.
- Regression-based approach: The regression based approach models the determinants of outcomes and possibly also models the determinants themselves. 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.⁹ 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.

Qualitative (participatory) impact evaluation¹⁰

Participatory approaches are intended to empower beneficiaries by enabling them to shape decisions which affect their lives (see Roche, 1999, for an exposition of participatory approaches to impact assessment). Participatory research eschews the pre-set agenda implicit in the structured questionnaires used in the quantitative approach. Rather techniques developed for Participatory Rural Analysis (PRA), such as village mappings, wealth rankings and oral life histories, are used. Through various means, beneficiaries comment directly on how the intervention has affected their lives (also called beneficiary assessment).

Projects with a participatory orientation are likely to have participation in the design of their M&E system, which means that beneficiaries played a role in defining the indicators to be monitored and perhaps also in the monitoring.

The advantage of qualitative approaches is that they usually provide a much stronger context and so a good ‘feel’ for the intervention. Stories of people’s lives and children’s drawings are more appealing to many readers than tables of statistical significance and scatter plots. However, participatory methods cannot deliver on rigorous impact evaluation: that is identifying attribution for changes in development outcomes. Critics would be even more severe, arguing that the samples used are not representative, so that there are no reliable data on outcomes amongst the treatment group even if these are measured – and there is no control so attribution cannot be estimated.

Participatory impact evaluation will not fill the evaluation gap as currently identified. But that does not mean it should be discarded. Participatory approaches are very likely to be the most suitable for process aspects of impact evaluations, and of course for process evaluations. Participatory evaluations also play an important role in furthering organizational change in institutions and projects as it is an inclusive stakeholder consultation which involves ownership of the assessment and facilitates the application of

9. When a dummy variable is used for project participation then the regression approach and PSM will give identical results if the same set of X variables is used in both cases.

10. The identification of qualitative with participatory needs to be treated with caution. Not all qualitative methods are participatory, and some participatory methods can give quantifiable data.

changes needed to be made. The best approach to impact evaluation combines quantitative and qualitative methods. Theory-based approaches tend to do just that.

Theory-based impact evaluation

Theory-based impact evaluation (TBIE) is not an alternative to quantitative or qualitative approaches. Rather it is a method for evaluation design.

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 provided by the intervention are expected to affect the outcomes specified in the objectives. In many cases this analysis will already be contained in the project log frame, which 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 – though usually supplemented by qualitative data.

A theory-based approach critically 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. There may also be weak links in the chain as a result of poor implementation. The links are analyzed through a combination of quantitative and qualitative argument: in some cases statistical analysis may be used to confirm a link, but in others a more qualitative case for plausible association may be made.

Building Support for Impact Evaluation

Bilateral agencies undertake their own impact evaluation

Evaluation departments can, and do, commission their own impact evaluation. However, the cost of such studies rules this out altogether for some agencies, or make it an infrequent occurrence for others. Since in most agencies there are more evaluations done outside of the evaluation department than inside, then the department may also play a role as a catalyst for rigorous impact evaluation. This approach is being taken by the DIME initiative of the World Bank's research department (and the broader Task Force on Impact Evaluation). The initiative supports a lunch-time seminar series and maintains a database of studies and consultants. The initiative has also fostered sector-specific events, usually workshops on IE design.

The advantage of this approach is that it allows for prospective evaluation, i.e. ensuring that there is a good evaluation design in place from the start, rather than having to find a way of creating a control group *ex post*.

Evaluation Departments may be wary of involvement with operational staff as it may be seen to compromise their independence. IEG, which is planning to undertake prospective impact evaluations, has adopted the position that they can advise on evaluation design but not project design. This approach is intended to keep evaluation staff at arm's length from operations.

Bilateral agencies can co-finance studies on programs of common interest

Given both the cost of IE studies and the increasing coordination through sector programs, donors may wish to co-finance a study on an area of common interest. Even if no sector program exists, it may still be possible to undertake a study of a particular sector in a country of interest to a group of donors.¹¹

The clear advantage of this approach is that it shares costs, whilst making policy relevant information available to all. The drawbacks are the well known drawback of multi-donor studies, i.e. coordination difficulties in agreeing terms of reference, who to commission to do the study etc. The most practical way of handling this problem is for one donor to act as the manager of the study, with the other agencies being sleeping partners. If the same group were to undertake a small number of studies management could rotate.

Bilateral agencies contribute to a common pool for impact evaluation

Since the outputs from impact evaluation have a large public good element, there is a case to be made for all donors paying a levy (though this would be voluntary) supporting a fund which finances IE. This is proposal made by the Center for Global Development for an Impact Evaluation Club.

The advantage of this approach is that it costs each donor less. However the approach may have a number of disadvantages. One is that, given the diversity of approaches, it would be a mistake to create a monopoly in the production of impact evaluations. Second, agencies will have less direct control over content and so are likely to be less able to ensure lesson learning from the studies. A network is a more appropriate concept than a club, moreover it should preferably be a network which has some roots in developing countries and does not have an undue orientation toward US-based institutions.

Issues for discussion by DAC Evaluation Network

Since a network is the desired way forward, it seems clear that there should be a substantial role of the existing evaluation network, i.e. the DAC Evaluation Network, though the Evaluation Cooperation Group (ECG) may also have a role. An acknowledged drawback on the part of many DAC members is their relative lack of experience in carrying out rigorous impact evaluation. IEG, with its long association with the DAC Evaluation Network, is in a good position to lend the necessary technical support.

There are several possible specific roles for the DAC Evaluation Network:

1. A facilitator of IE studies by individual or groups of bilateral donors. The Network has in the past facilitated the launch of many joint evaluation initiatives and could similarly be the focal point for the formation of “evaluation coalitions” of members willing to move forward and take part in and sponsor IE work.
2. The Network could support the development of IE materials to assist individual agencies in undertaking IE. However, such materials are to some extent already available (e.g. Baker, 2000) or under development in the World Bank and elsewhere. There may still be scope for a lesser role in organizing events in support of IE. This may be done in collaboration with IEG (who may also draw upon other resources in the World Bank).
3. Acting as a partner or host for the proposed Impact Evaluation Club, depending on what will come out of the proposals contained in the CDG report. However, as noted above, a more broadly

11. Though the challenge to the evaluation design should not be under-estimated.

conceived network appears more appropriate than a club (broader both methodologically and in membership).

4. Acting as a clearing house to distribute findings from impact evaluations. IEs are a public good only if the findings from them are well-known, whereas agencies do not always share results. The DAC Evaluation Network could ensure that not only is an inventory kept of IEs (which the World Bank is already doing though not in a comprehensive manner), but also a way to disseminate results in an accessible manner.
5. As a DAC body, the Evaluation Network can facilitate the feed back findings and lessons to the relevant policy communities in development agencies. Any launch of new studies will need to be accompanied by measures to actively support lessons learning on what works.

Summary

The main points in this note:

1. There is a need for more impact evaluation, both to demonstrate results and for lesson learning.
2. There are a variety of approaches, but rigorous impact evaluation requires uses of quantitative methods.
3. There is scope for greater use of experimental (randomized) evaluation design, and program managers need be aware of this possibility.
4. But randomized approaches are not applicable in a great number of cases, so that other methods need also be employed.
5. Good impact evaluation also pays attention to context and strives for policy-relevance.
6. Any proposed program of impact evaluation should be based on a broad partnership, including participants from developing countries.
7. There is an important role for the DAC Evaluation Network in helping bring such a program forward. IEG can provide technical support to these efforts.

Annex: Quantitative impact evaluation¹²

The cornerstone of quantitative impact evaluation is data collection from a statistically representative sample using a structured questionnaire. It is strongly preferable that data are collected both before the intervention (baseline) and after (endline). A midterm survey is also an advantage. Data should be collected from both the affected population (the treatment group) and a comparison group.¹³ Project impact is then calculated as either a single difference (difference in outcome between project and control after the intervention), or double difference (the difference in the change in outcomes for the project and control before and after the intervention).

Selection of an appropriate comparison group is one of the main challenges in impact evaluation. They should be identical to the treatment group except that the latter receive the intervention and the former do not. In practice this is difficult to achieve for two reasons. First, beneficiaries of the intervention may be selected (or self-select) on the basis of certain characteristics. If these characteristics are observed then a comparison group with the same characteristics can be selected. But if they are unobserved then in principle only a randomized approach can eliminate selection bias.¹⁴ Second, the comparison group may be contaminated either by spillover effects from the intervention or a similar intervention being undertaken in the comparison area by another agency.

Experimental (randomized) approaches: Experimental (or randomized) evaluation design tackles the first, but not the second, of these problems. Experimental design requires that the eligible population be identified and then a random sample of those treated. 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.

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 project, 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.

However, there are 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, since discrete, homogenous interventions – like taking a pill – are most

12. This Annex is a longer version of the subsection on quantitative impact evaluation.

13. Usually referred to as the control group. Strictly speaking this term is only correct in an experimental design in which after factors are controlled so as not to vary between treatment and control.

14. This statement needs to be qualified in two ways. First if the unobservables are time invariant they can be removed by double differencing. Second, if the direction of the bias can be reasonably assumed then upper or lower estimates of project impact can be made.

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. 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. 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 development agencies. 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.

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.

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. The difference in the mean outcome from the two groups is taken as project impact.¹⁵

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.

The potential problem with PSM is that facing all quasi-experimental approaches: selection on unobservables. Unobservables which simply affect project outcomes and 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.

15. 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 models the determinants of outcomes and possibly also models the determinants themselves. 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.¹⁶ 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.

16. When a dummy variable is used for project participation then the regression approach and PSM will give identical results if the same set of X variables is used in both cases.

References

Baker, Judy (2000) *Evaluating the Poverty Impact of Projects* World Bank. (http://imagebank.worldbank.org/servlet/WDSContentServer/IW3P/IB/2000/08/19/000094946_00080705302127/Rendered/PDF/multi_page.pdf)

Center for Global Development (2005) 'When will we ever learn? Recommendations to improve social development through enhanced impact evaluations', September 2005 (http://www.cgdev.org/section/initiatives/_active/evalgap)

OED (2005) 'OED and Impact Evaluations – a discussion note' (http://www.worldbank.org/oed/docs/world_bank_oed_impact_evaluations.pdf)

Roche, Chris (1999) *Impact Assessment for Development Agencies: learning to value change*. Oxford: Oxfam.