

OUTLINE OF PRINCIPLES OF IMPACT EVALUATION

PART I KEY CONCEPTS

Definition

Impact evaluation is an assessment of how the intervention being evaluated affects outcomes, whether these effects are intended or unintended. The proper analysis of impact requires a counterfactual of what those outcomes would have been in the absence of the intervention.¹

There is an important distinction between monitoring outcomes, which is a description of the factual, and utilizing the counterfactual to attribute observed outcomes to the intervention. The IFAD impact evaluation guidelines accordingly define impact as the “the attainment of development goals of the project or program, or rather the contributions to their attainment.” The ADB guidelines state the same point as follows: “project impact evaluation establishes whether the intervention had a welfare effect on individuals, households, and communities, and whether this effect can be attributed to the concerned intervention”.

The counterfactual

Counterfactual analysis is also called with versus without (see Annex A for a glossary). This is not the same as before versus after, as the situation before may differ in respects other than the intervention. There are, however, some cases in which before versus after is sufficient to establish impact, this being cases in which no other factor could plausibly have caused any observed change in outcomes (e.g. reductions in time spent fetching water following the installation of water pumps).

The most common counterfactual is to use a comparison group. The difference in outcomes between the beneficiaries of the intervention (the treatment group) and the comparison group, is a single difference measure of impact. This measure can suffer from various problems, so that a double difference, comparing the difference in the change in the outcome for treatment and comparison groups, is to be preferred.

Purpose of impact evaluation

Impact evaluation serves both objectives of evaluation: lesson-learning and accountability.²

A properly designed impact evaluation can answer the question of whether the program is working or not, and hence assist in decisions about scaling up. However, care must be taken about generalizing from a specific context. A well-designed

¹ “The central objective of quantitative impact evaluation is to estimate... unobserved counterfactual outcomes” *Impact Evaluation: methodological and operational issues*, Asian Development Bank, September 2006; and “determining the counterfactual is at the core of evaluation design”, Judy Baker *Evaluating the Impact of Development Projects on Poverty*, World Bank, 2000.

² See the DAC Evaluation Network report, *Evaluation Feedback for Effective Learning and Accountability Report No 5, OECD Evaluation and Effectiveness Series*, OECD (2001) for a fuller discussion of these functions of evaluation.

impact evaluation can also answer questions about program design: which bits work and which bits don't, and so provide policy-relevant information for redesign and the design of future programs. We want to know why and how a program works, not just if it does.

By identifying if development assistance is working or not, impact evaluation is also serving the accountability function. Hence impact evaluation is aligned with results-based management³ and monitoring the contribution of development assistance toward meeting the Millennium Development Goals.

When to do an impact evaluation

It is not feasible to conduct impact evaluations for all interventions. The need is to build a strong evidence base for all sectors in a variety of contexts to provide guidance for policy-makers.

The following are examples of the types of intervention when impact evaluation would be useful:

- Innovative schemes
- Pilot programs which are due to be substantially scaled up
- Interventions for which there is scant solid evidence of impact in the given context
- A selection of other interventions across an agency's portfolio on an occasional basis

PART II EVALUATION DESIGN

Key elements in evaluation design

The following are the key elements in designing an impact evaluation:⁴

- Deciding whether to proceed with the evaluation
- Identifying key evaluation questions
- The evaluation design should be embedded in the program theory
- The comparison group must serve as the basis for a credible counterfactual, addressing issues of selection bias (the comparison group is drawn from a different population than the treatment group) and contagion (the comparison group is affected by the intervention or a similar intervention by another agency).
- Findings should be triangulated
- The evaluation must be well contextualised

³ See the report for the DAC Evaluation Network *Results-Based Management in Development Co-Operation Agencies: a review of experience* (2000), for a discussion of the relationship between evaluation and results-based management (RBM). The UNEG Task Force on Evaluation and Results Based Management is currently examining this relationship in further detail.

⁴ This list is adapted from Baker op. cit.

Establishing the program theory

The program theory documents the causal (or results) chain from inputs to outcomes.⁵ The theory is an expression of the log frame, but with a more explicit analysis of the assumptions underlying the theory. Alternative causal paths may also be identified. The theory must also allow for the major external factors influencing outcomes.

A theory-based evaluation design tests the validity of the assumptions. The various links in the chain are analyzed using a variety of methods, building up an argument as to whether the theory has been realized in practice.

Using the theory-based approach avoids 'black box' impact evaluations. Black box evaluations are those which give a finding on impact, but no indication as to why the intervention is or is not doing. Answering the why question requires looking inside the box, or along the results chain.

Selecting the evaluation approach

A major concern in selecting the evaluation approach is the way in which the problem of selection bias will be addressed. How this will be done depends on an understanding of how such biases may be generated, which requires a good understanding of how the beneficiaries are identified by the program.

Figure 1 (Annex B) shows a decision tree for selecting an evaluation approach. The basic steps in this decision tree are as follows:⁶

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

⁵ Chapter 3 of Carol Weiss (1998), *Evaluation*, gives an exposition of program theory. Both the IFAD and AfDB impact evaluation guidelines strongly espouse a theory-based approach.

⁶ The various approaches are discussed in a number of sources, including the ADB booklet, *Impact Evaluation*, Baker op cit, and Martin Ravallion (1999) *The Mystery of the Vanishing Benefits: Ms Speedy Analyst's Introduction to Evaluation*, World Bank Policy Research Working Paper 2153.

Designing the baseline survey

Ideally a baseline survey will be available so that double difference estimates can be made. Important principles in designing the survey are:

- Conduct the baseline survey as early as possible.
- The survey design must be based on the evaluation design which is, in turn, based on the program theory. Data must be collected across the results chain, not just on outcomes.
- The comparison group sample must be of adequate size, and subject to the same, or virtually the same, questionnaire. Whilst some intervention-specific questions may not be appropriate, similar questions of a more general nature can help test for contagion.
- Multiple instruments (e.g. household and facility level) are usually desirable, and must be coded in such a way that they can be linked.
- Survey design takes time. Allow six months from beginning design to going to the field, though 3-4 months can be possible. Test, test and re-test the instruments. Run planned tabulations and analyses with dummy data or the data from the pilot. Once data are collected one to two months are required for data entry and cleaning.⁷
- Include information to allow tracing of the respondents for later rounds of the survey, and ensure that they can be linked in the data.
- Avoid changes in survey design between rounds. Ideally the same team will conduct all rounds of the survey.

Options when there is no baseline

Evaluations are often conducted ex post, and there is no baseline available. Under these circumstances the following options can be considered.⁸

1. If treatment and comparison groups are drawn from the same population and some means is found to address selection bias (which will have to be quasi-experimental, since randomization is ruled out unless the treatment had been randomized, but if the program designers had thought of that they will have thought of a baseline also), then a single difference estimate is in principle valid.
2. Find another data set to serve as a baseline. If there was a baseline survey but with a poor or absent comparison group, then a national survey might be used to create a comparison group using propensity score matching.
3. Field a survey using recall on the variables of interest. Many commentators are critical of relying on recall. But all survey questions are recall, so it is a question of degree. The evaluator need use his or her judgment as to what it is reasonable to expect a respondent to remember. It is reasonable to expect people to recall major life changes, introduction of new farming methods or crops, acquisition of large assets and so on. But not the exact amounts and prices of transactions. When people do recall there may be telescoping (thinking things were more recent than

⁷ Baker, op cit., allows nine months from survey design to obtaining cleaned data.

⁸ Partly based on the discussion in the IFAD impact evaluation guidelines.

they were), so it is useful to refer to some widely known event as a time benchmark for recall questions.

4. If all the above fail, then the study was make build a strong analysis of the causal chain (program theory). Often a relatively descriptive analysis can identify breaks in the chain and so very plausibly argue there was low impact.
5. The argument can be further strengthened by triangulation (indeed this point applies whatever method is adopted): drawing on a variety of data sources and approaches to confirm that a similar result obtains from each.

Impact evaluation using secondary data

Sometimes secondary data can be used to carry out the whole impact study,⁹ this is especially true when evaluating national or sector-wide interventions. More usually secondary data can be used to buttress other data. For example, a project data set could be used for the treatment group and a national data set used to establish the control, preferably using a matching method. Or different national data sets might be joined to enable a rigorous regression-based approach to be employed.

The role of qualitative information

Good evaluations are almost invariably mixed method evaluations¹⁰. Qualitative information informs both the design and interpretation of quantitative data. In a theory-based approach, qualitative data provide vital context.

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 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. 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 should not be overlooked.. 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. Return trips are also advisable to help elucidate the findings.

Triangulation

Evaluation findings are strengthened when several pieces of evidence point in the same direction. Often a single data set will allow a variety of impact assessments to be made. Better still if different data sets and approaches can be used and come to broadly the same conclusion. Qualitative information can also reinforce findings and add depth to them. Where a rigorous approach has not been possible then

⁹ For further discussion see Michael Bamberger *Conducting Quality Impact Evaluations under Time and Budget Constraints*, 2006, World Bank.

¹⁰ For further discussion see Michael Bamberger (2000) *Integrating Quantitative and Qualitative Methods in Development Research*, World Bank and Weiss op. cit.

triangulation is all the more necessary to build a case based on plausible association.¹¹

Generalisation from specific impact evaluations

Impact evaluations are usually of specific interventions in a specific context. It is not necessarily the case that the findings can be generalized to the same intervention in different contexts. A theory-based approach helps understand the context in which the intervention did or didn't work, and so help generalize as to other contexts in which the same findings may be expected.

PART III MANAGING AND IMPLEMENTING IMPACT EVALUATION

Terms of reference

The ToR should require that a clear understanding of the intervention be a prerequisite for the evaluation design. Sector and area expertise may not be essential but is certainly an advantage.

The ToR for an impact evaluation should also stress the need for a credible counterfactual analysis. Proposals or concept notes should make clear how this issue will be addressed, being explicit about the evaluation approach. The evaluation team need include personnel with the technical competence to implement these methods.

Data sources

Good quality data are essential to good impact evaluation. The evaluation design must be clear on the sources of data and realistic about how long it will take to collect and analyze primary data.

Time and cost

The time required for an impact evaluation depends on whether primary data collection is involved. If so, 18 months is a reasonable estimate from inception to final report. If there is no primary data collection then 12 months may be feasible.

The survey is largest cost component of an impact evaluation. If primary data are being collected then the total cost of a survey is approximately US\$ 100 per household in Africa and Latin America, US\$ 40-60 in East Asia and US\$ 25-40 in South Asia. Analysis time is 3-4 months of professional level inputs.

Peer review

An independent peer review should be undertaken by a person qualified in impact evaluation.

¹¹ This is the approach advocated in the AfDB impact evaluation guidelines.

ANNEX A GLOSSARY

Note: glossary follows a natural sequence rather than alphabetical order

Attribution: The problem of attribution is the problem of assigning observed changes in outputs and outcomes to the intervention. This is done by constructing a counterfactual.

Counterfactual: outputs and outcomes in the absence of the intervention. The counterfactual is necessary for comparing actual outputs and outcomes to what they would have been in the absence of the intervention, i.e. with versus without.

With versus without: 'with' refers to the outputs and outcomes with the intervention (*the factual*), which are compared with the outputs and outcomes 'without' the intervention (the counterfactual) to determine the impact of the intervention, though single or double difference estimates.

Comparison group: For project-level interventions, the counterfactual is often established by taking a comparison group (typically a geographic area) which is identical to the treatment group, except that it is not subject to the intervention. (The expression 'control group' is also commonly used, but strictly speaking only applies to experimental settings in which the conditions of the control group can be controlled).

Experimental design: In order to ensure comparability, an experimental design randomly assigns eligible households to the project and comparison groups. This approach can avoid selection bias, but the extent to which it can be applied to the types of intervention supported by DFID has been questioned.

Selection bias: The beneficiaries of an intervention may be selected by some criteria (or select themselves) which is correlated with the observed outcome. For example, entrepreneurs being selected for microcredit or for a business development scheme may have done better than those who did not bother to apply, even in the absence of the support. Hence comparing outcomes of beneficiaries and non-beneficiaries can give misleading results. Where these criteria are not observed (i.e. there are no data on them), then there is a bias in the impact evaluation findings (this point is discussed further below). But where the determinants of participation are observed, then the bias can be removed using quasi-experimental methods.

Quasi-experimental design: evaluation designs which address selection bias using statistical methods, such as propensity score matching, rather than randomization. These methods model the selection process and so control for these variables in the analysis of outcomes.

Contagion or contamination: The comparison group is contaminated if it is subject to a similar intervention, either by spill-over effects from the intervention or another donor starting a similar project.

Single difference: the difference in the output or outcome either (1) before versus after the intervention, or (2) between project and comparison groups. Before versus after is not a good impact measure as it fails to control for other factors. The single difference project versus comparison groups fails to allow for differences between the two groups which may have existed prior to the intervention. The double difference takes care of these two problems.

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



