Evidence-based Evaluation of Development Cooperation: Possible? Feasible? Desirable?

Kim Forss and Sara Bandstein

1 Introduction and purpose

International development cooperation in recent years has seen an increased interest in evidence and evidence-based policies and practices. The central idea of the evidence movement is for policies and practices to be based on the best available scientific research about what works, what does not, and the reasons why. However, the evidence notion is by no means unambiguous and what constitutes evidence is highly debatable. In this article, evaluation is defined as evidence-based provided it makes explicit use of a counterfactual. This can be done through experimental or quasi-experimental research designs.

Evidence-based evaluations have appeared later in international development cooperation than in areas such as medicine, social work, and education. The lack of impact evaluations has been increasingly recognised.[1] Many researchers and evaluators, and indeed administrators, managers and politicians, call for and look for evidence of results of development spending. But a brief review of evaluation practice suggests that such evidence is not forthcoming, despite the significant amounts of money spent on evaluation. This article sets out to explore why, and discuss how to improve the situation. It has three interrelated purposes.

1 *Descriptive:* given a strict definition of what constitutes a sound and reliable design for evidence-based evaluation, how often do evaluations live up to such standards?[2]

2 *Explanatory:* why do so few evaluations use experimental or quasi-experimental enquiry methods? We focus on how evaluation processes are determined and what role terms of reference (ToRs) play in deciding research design.

3 Assuming more evidence-based evaluations would be desirable, our third purpose is to produce hypotheses on how to generate more of these, to discuss what changes this necessitates among those who commission and use evaluations, and among those who conduct them and/or are subjected to them.

2 What do evaluations evaluate?

To assess the extent to which evaluations of development cooperation use experimental or quasi-experimental design, a survey was made of 80 evaluations conducted in the period 2004–7. These were randomly selected from the Development Assistance Committee (DAC) Evaluation Resource Centre, DEReC, providing evaluations commissioned by member organisations of the DAC evaluation network. To date these organisations have submitted over 1,500 reports to DEReC.

For the years selected for this study, 2004–7, 345 evaluation reports were available in the database, of which our random sample constituted 23 per cent of the total. The overall picture of aid evaluation we get from these 80 evaluations should reflect the state of the art as practised by official agencies. The 80 reports were commissioned by 22 organisations, five of which were commissioned jointly by two or more organisations or partner countries.

Our evaluation sample consists mostly of project and programme evaluations. Over one-third are project evaluations and almost as many assess programmes. Of the latter, most evaluate country or regional programmes. In one-fifth of the reports, organisations, methods or policies and strategies are the main objects of inquiry. Although it is claimed

that the overall trend is for aid to become more sector- and policy-oriented, many of the evaluations conducted in the years 2004–7 still examine projects.

A relatively large number of the reports (10 per cent) cannot be classified as evaluations at all, being annual reports, policy documents, background papers or overviews; short summaries without main reports or management audits.

The choice of research design is highly dependent on the timing of the evaluation. Our assessment of the 80 evaluations shows that almost two-thirds of the evaluated activities were ongoing when the evaluation was carried out. Many of these were country programme or sector evaluations, some projects being ongoing and some completed. Less than one-third of the reports assess completed activities. Almost all of those which did so were project evaluations. However, in most cases the evaluation was undertaken shortly or immediately after the activity was completed. Only in rare cases were the evaluations carried out several years after the activities had ended.

2.1 Evaluation criteria and research designs

Apart from reports classified as non-evaluations, all studies assess effects of some kind: an intervention, a policy, a method or an organisation. DAC evaluation criteria of relevance, effectiveness, efficiency, impact and sustainability were largely used in the evaluations. However, we did not analyse in depth whether the criteria were actually used, whether they were used correctly, or if they were only mentioned. If the evaluation claimed to use a certain criterion, it was accepted. Hence, of the evaluations, two-thirds claimed to assess effectiveness and relevance of interventions and over half the reports used the criteria of impact, efficiency, and sustainability.

The research designs of the evaluations are largely homogenous. There was only one case of truly experimental design. This assessed the impacts of the development of genetically improved farmed tilapia (GIFT) and their dissemination in selected countries and was commissioned by the Asian Development Bank. In a few cases, (5.2 per cent) a counterfactual was created through quasi-experimental research design. Of these, only one evaluation explicitly addressed the issues of selection bias and spillover effects. Taken together, less than 7 per cent of the evaluations explicitly make use of counterfactual analysis. Most evaluations are conducted non-experimentally.

2.2 Methods

Various methods were used to collect evaluation data. Most evaluations (nine out of ten) used documentary analysis and interviews. In half the samples, observations or longer field trips were used as data sources. Fifteen per cent of the cases gained information through questionnaires. Surprisingly, the different kinds of data collection methods are not correlated to the different research designs. Baseline data are, for instance, used to the same extent in quasi-experimental studies as in non-experimental. Questionnaires, however, seem to be used more frequently for evaluations with a quasi-experimental design. A full statistical analysis was not possible due to the small number of experimental and quasi-experimental studies.

To conclude, the range of methods is limited and few reports enable a critical reader to form an opinion on the reliability of the study findings. Analyses of how documents are selected for analysis, or used and assessed, are rare. Evaluations including information on selection of interview respondents or interview methods are also rare. Only in a few cases are surveys and questionnaires attached to the evaluation report.

3 The demand for information – the terms of reference

We must remember that the situation described above is the result of purposeful and rational action. It reflects decisions taken by management in bilateral and multilateral funding agencies. The demand for evaluative information is expressed in the terms of reference (ToRs) for the evaluations, so if we look for reasons why evaluations produce – or don't produce – any evidence, we should start by looking at the ToRs.

It is not uncommon for evaluations to fail to respond to their ToR, but when they do so, there is usually a negotiation between the evaluators/consultants and the client. The latter could, in theory, refuse to accept a product that does not correspond to the ToR. The development agency could request the evaluators to finalise the inquiry and provide the evidence requested. Furthermore, if the evaluation process had failed, which they sometimes do, the aid

Table 1 Demand for evaluative information in ToR

Included in the ToR	Efficiency	Effectiveness	Impact	Sustainability	Relevance
Yes	53	61	56	51	66
No	8	28	17	31	29

agency would not need to publish the report, or include it in the OECD/DAC database. Hence we may assume that the evaluations in our sample are studies that the organisations were satisfied with and that have responded adequately to their ToRs.

3.1 Results

All ToRs asked for information about results. However, the notion of 'results' is not as straightforward as it sounds. Collins English Dictionary (1982) says that a result is 'something that ensues from an action, policy, etc.; outcome, consequence'. A 'result' is also synonymous with an 'effect'. The Evaluation Thesaurus (Scriven 1991) says that 'Effect' is 'An outcome or type of outcome'. 'Outcome' in the same Thesaurus is defined as 'post treatment effects.' The circle has been closed, effects are outcomes and outcomes are effects, and both are results, but are we any the wiser?

The DAC glossary of key terms in evaluation and results-based management defines 'Results' as 'The outputs, outcomes and impacts (intended or unintended, positive and/or negative) of a development intervention' (2002). This goes back to the common-sense use of the term implied in the Collins Dictionary; analysis of results concerns what has happened after an intervention, and can be in the long- or in the short-term, more or less directly caused by the intervention.

It is perhaps not necessary to make things more complicated than they already are. Those who commission evaluations may know exactly what they ask for when evaluators are commissioned to analyse results. However, what to look for and how to look for it would certainly vary significantly across contexts as different as, for instance, budget support from the international donor community to a wide number of developing countries over many years, or World Bank lending to China over the past decade, or Finnish development cooperation in the education sector, or a three-year project to strengthen local radio stations in Vietnam.

Hence when the ToRs ask the evaluators to document results, this request must be understood in the very specific context of an intervention, how long it has been going on, what it tried to achieve, and what other factors influence events in that context. The word 'result' itself has little meaning but must be elaborated carefully in the context of the intervention.

3.2 Dimensions of results analysis

It is often qualified in the ToRs that results should be analysed against the OECD/DAC criteria of efficiency, effectiveness, impact, sustainability and relevance. These are all related to results; you could even say that they provide dimensions of an analysis of results. Table 1 indicates to what extent the evaluations were asked to analyse results in one or several of these dimensions. None of the ToRs ask for information about only one aspect of results and the majority ask for information about all five aspects of results. What then are the consequences for the design of evaluations?

- First, an analysis of *efficiency* does not require an experimental or quasi-experimental design. When evaluators analyse efficiency they need to understand how the results were produced or the costs of an intervention. That cannot be done by experimental design. The concluding evidence of efficiency is usually based on a benchmark, rather than on data from another intervention or a comparison group.

- Second, an analysis of *sustainability* takes the evaluator into the future and hence experimental design cannot prove any better evidence than any other method. Here the evaluator uses hypothetical arguments, speculates about what might happen, and analyses the conditions for sustainable effects; this is not achieved through experimental design.

- Third, *relevance* cannot be assessed through experimental or quasi-experimental design either. If the intervention in respect of the experiment

group is relevant, then the intervention with the comparison group is equally relevant – or irrelevant.

- Fourth, while it is relatively easy to see how experimental or quasi-experimental designs could be used to analyse *effectiveness* and *impact*, it is less clear whether both could be done within the same research design. The analysis of effectiveness would in most instances require its own set of experiment and comparison groups, depending on the population, the duration of interventions, etc. But the analysis of impact goes further and would normally involve other groups in society; hence the experiment and comparison groups would differ.

While the four points above are purely theoretical and relate to the dimension of results as such, it is also useful to consider the practical aspects of the evaluation – namely what is it that is being evaluated and what the consequences for the prospects of experimental and quasi-experimental research design are.

Although many of the evaluations assessed projects, almost as many were directed at a sector or a country. If the evaluation is asked to look for results in respect of a five-year programme of development cooperation in, for example, Lao PDR, what are the design options? Assume that such an evaluation would focus on effectiveness only. Where is the experiment group and where is the comparison group? In a sector programme many donor agencies together contribute to a common basket to fund a large number of different activities in, for example, the health sector. The activities can range from the purchase of drugs to the training of doctors and building hospitals and health clinics, and institutional strengthening to manage the sector. Again, where are the experiment group and the comparison group, and could a government be expected to plan its use of foreign resources in the form of experiment and comparison groups in the health sector?

Looking at the cases where in theory it would be possible to use an experimental approach, that is, where the evaluated intervention is a project or a similarly well-defined unit, why are such designs not chosen more often? The answer may lie in the multitude of dimensions of results. If the evaluators are asked to analyse all five dimensions, and three of them cannot be answered through experimental methods, and the other two require two different designs, then perhaps the chances are that the resources are not sufficient, or that in a comparative light, these two questions are not that important. Perhaps many consider it more worthwhile to spend resources on a valid analysis of relevance and sustainability than on an experimental design to analyse effectiveness and impact?

3.3 Aspects of management

But not only do the ToRs ask many questions about results, they also ask about how the results were produced in terms of various aspects of management and implementation (92 per cent of the cases), for example;

> Describe the governing structure between MFA, Fredskorpset, the partner organisations and the individual participants, the internal management structure of Fredskorpset, the internal division of labour, and the available management tools to administrate the division of labour. Describe categories of challenges participants have encountered during their stay, the monitoring mechanisms to capture such problems, and responsibilities to support the participants when possible problems arise. Assess the strengths and weaknesses of today's governing structure and division of labour.
> (Norad, *Evaluation Report* 2/2006, Evaluation of Fredskorpset)

The development banks often have very similar ToRs for their country evaluations. One typical example is the evaluation of the World Bank's country programme in Armenia in which the evaluators were asked to analyse the design of the country programme, as well as planning and management. Since it is often not clear what is meant by 'management', different evaluators will choose to look for and comment on different things. Some analyse the organisational structures put in place to implement the programme; they may look at decision-making structures, for example the decentralisation or centralisation of decision-making powers. Others look at the comparison of funds, the presence of audit reports, the prospects for corruption. Yet others discuss coordination mechanisms, and the level or shortcomings of coordination; for example within the country programme, between the bank and other donors,

between the bank and the government, or between headquarters and field office. Whatever focus the evaluators choose, they usually penetrate issues through interviews and through an analysis of the managerial issues based on a theoretical and practical understanding of what the problems could be and how they can be resolved.

Is it at all possible to choose an experimental or quasi-experimental design to generate evaluative information on management and implementation of the interventions? Research in the managerial sciences, organisation theory, administrative behaviour, etc. frequently makes use of experimental and quasi-experimental designs. A quick review of scientific journals such as *Administrative Sciences Quarterly*, *Organisation*, or *Journal of International Business Studies* proves the point. The research frontier is moved forward through a combination of qualitative case study research as well as quantitatively-based experimental research. So if research questions can be answered with the help of experimental methods, why cannot evaluation questions be answered the same way? To answer that let us look at the questions asked:

> [the evaluation should] Assess the Programme's developmental performance to date (including the Programme's cumulative performance since 1994) with respect to results achievement, sustainability and relevance, as well as its operational performance regarding partnerships, informed and timely action and resource utilisation; Identify key issues emerging from reviews of the context of CCPP stakeholders – Canadian International Development Agency (CIDA), other donors, Canadian College Institutions (CCIs), Developing Country Organisations (DCOs) and Developing Countries (DC) – that may positively or negatively affect the relevance of CCPP.
> (Evaluation under the Canadian colleges partnership programme phase II 2001–8)

Another example:

> [In the course of the evaluation] particular attention should be paid to the following aspects:
>
> - Project outputs and impact, as far as feasible, compared with project design;
> - The institutional arrangements of the project, both operational and managerial;
> - The coordination between the project and the Chinese authorities, national, provincial and municipal and project consultants and institutions;
> - The relative performance of the project and its institutions in the different provinces and municipalities;
> - The performance and coordination of the consultants' consortium;
> - DFID's arrangements for monitoring the project and liaising with the Chinese authorities;
> - The impact and effectiveness of the consultants' output-based contractual arrangement;
> - The extent to which it can be shown that the project's examples were replicated elsewhere in the pilot municipalities and provinces, or more widely.
> (*DFID Evaluation Report* EV 658. Review of the China State-Owned Enterprise Restructuring and Enterprise Development Project)

We are not suggesting that these questions are not interesting and relevant, or that answers would not be useful in the future management of the activities. But the research design that, in a few months' time, could provide answers, would certainly not be experimental or quasi-experimental.

3.4 Purpose of the evaluations

If these are the questions asked (about results, efficiency, effectiveness, impact, sustainability, relevance, and a number of aspects of management and implementation), why are they asked? This is a trivial question and the answer is also trivial. Because all of the agencies emphasise that they are results-oriented, they manage for results and they apply, since many years back, some form of results-based management system. Hence they need information on results in order to take decisions on future interventions. Furthermore, no more than 24 of the interventions evaluated had been completed, while 45 were still going on or were proceeding into a new phase (in 11 cases we could not tell whether the intervention had come to an end, or the question was irrelevant).

Therefore, the choice of evaluation design should also be seen in the light of its purpose. Table 2 presents a summary of evaluation purposes. We have only looked at the explicit purposes mentioned in the ToRs and not assessed ritualistic elements in evaluation purpose nor other 'illegitimate' uses of evaluation that are frequently mentioned in evaluation theory (DsUD 1990: 63; Vedung 1997).

Table 2 The purpose of evaluations according to ToRs*

Purpose	Mentioned in number of ToRs
Comparison	15
Decision support	28
Learning	19
Total that define one or more purposes	40

* in 10 of the evaluations/ToRs it was not possible to deduct any purpose beyond getting the information, that is, nothing concerning how it was to be used.

Most evaluations are undertaken for a managerial purpose and are expected to provide analytical support for decisions on the intervention. It is likely that information on implementation is more immediately useful than information on results, and hence a priority in the practical work.

If we look at the two ToRs quoted above, from CIDA and DFID, it is immediately clear that management in these organisations would benefit from knowledge about coordination processes, stakeholder relations, consultancy contracts, etc. They could act on that information, implement changes within existing project and programme documents, modify contracts, terminate some activities and launch new ones. These are practical decisions where managers may find it useful to have some analytical support from the evaluations. But is it likely that an experimental or quasi-experimental study would provide that kind of information? No, not really; their strengths lie elsewhere.

4 The scope for evidence-based evaluation – what could be done

This article shows that evidence-based evaluation is not common in evaluation of development cooperation. We have also sought an explanation for why evidence-based evaluation is not more common, and found that the demand is simply not there. The purposes of the evaluations and the way in which evaluations are commissioned imply that it is highly unlikely that evaluators will choose an experimental or quasi-experimental design for their studies. In this section we turn to discuss how evaluation management may need to change if evidence-based evaluation is to become more frequent.

4.1 Focus the demand for information

There is a problem with the demand for information. It is not that the objectives of evaluation are unclear or that the questions are fuzzy. The problem is rather that there are too many questions to be answered at the same time. As we saw above, most of the evaluations are asked to analyse a number of complex managerial processes involved in how the interventions are implemented. They are also asked to analyse results in terms of efficiency, effectiveness, impact, sustainability and relevance. Hence, if evaluations are to increase the use of experimental methods, the task needs to be redefined.

As a first requirement it is necessary to distinguish evaluations that are done for managerial purposes, and so need information on management and implementation, as well as on relevance, efficiency and sustainability. Given that such issues are not best addressed through experimental studies, the risk is that they 'contaminate' the impact studies. The urgency of decision support and learning weighs heavier in the short run and determines the choice of evaluation design. When the aid agencies want impact information they should not confound those evaluation tasks with a number of other issues and questions, but should focus on impact and do nothing else.

Globally speaking, several thousand evaluations are commissioned by the various agencies in development cooperation. It is a major challenge to consider that most of these would have to be rethought, along with the interventions themselves. The most significant event in development cooperation in recent years was the signing of the Paris Declaration on aid effectiveness. This calls for national ownership, alignment, harmonisation, results-based management, and mutual accountability. While not inherently contradictory, it is not obvious that governments in developing countries would afford a high priority to experimental approaches to evaluation.

4.2 Time the evaluations

Not only would it be necessary to focus the assignments on questions that can be addressed through experimental methods – it would also be necessary to time the evaluations differently. Either the evaluations in our sample were commissioned in time for a mid-term review or some other decision

point in the course of an intervention, or once the intervention was completed. The experimental approach requires that the evaluation is designed when the intervention starts. And in its extreme form, with randomised experiment and comparison groups, it means that most common forms of development planning (which are highly political exercises) would have to be conducted according to an evaluation logic rather than in response to other demands, such as sector allocation, vulnerable groups, poverty orientation, etc.

4.3 Be specific about results

The word 'results' is often used as a mantra to solicit goodwill. It is always possible to assess some results, but it may not be possible to assess all results all the time. Most interventions covered in these 80 evaluations were very complex undertakings, and if they were successful they would have produced a large number of results, in the long term and in the short term, as well as a number of side effects. It is easy to commission an evaluation to document results, and most evaluations do. But if the evaluations are to be conducted with rigorous research methods they must not only be focused on results *per se*, they have to focus on one, or a few, particular results.

5 Concluding remarks

In this article we have taken seriously what in the evidence movement are said to be rigorous methods and sound evaluation practice. We have also taken seriously and described what happens in evaluation processes. There is a very wide gap between what many would consider desirable in terms of evidence of results and the practice of evaluation. Is it a gap that can be closed? We are pessimistic. There is little evidence that practice will change and in fact there are trends in development cooperation that make it less likely that development cooperation will be conducted in the form of policy experiments that may lend themselves to evidence-based evaluation approaches. But rather than confront that gap and the conflicts of interest inherent in it, decision-makers seem to turn to wishful thinking and hope that the problem will go away if the mantras of results-based management and evidence of results are repeated often enough.

We gave the title of this article three questions: whether evidence-based approaches are possible, feasible,[3] and desirable. It is time to suggest answers.

5.1 Possible: in most cases 'no'

Is it possible to provide evidence in response to the questions posed in ToRs as they are currently expressed? Most of the time it is not possible because:

1 The questions are too many; it is not possible to design a process of inquiry that could provide answers to the typical evaluation questions with an experimental design;

2 The timing is wrong and depends on the decision-making needs of the agencies rather than on the requirements to conduct reliable research;

3 The notion of results was not adequately defined. If there is not a precise and specific question to guide the research process it is not possible to design experimental studies.

This does not mean that it is impossible to get evaluative information on results, nor to get evaluative information on impact and effectiveness specifically. But the evaluative information will in all likelihood have to be based on case study analysis, narrative analyses, and the standards of research evidence common in many of the social sciences, for example in history, political science, economic history, anthropology, and sociology. Here evidence is not strictly associated with the experimentally conducted research. On the contrary, it is through qualitative studies and thorough application of rigorous and stringent research methods (but not through experimental design) that our knowledge of the social reality has grown.

5.2 Feasible: in most cases 'no'

Is it possible to provide evidence in response to the questions posed in ToRs, such as these are written today? Most of the time it is not feasible because:

1 In practice the willingness to commit large sums of money to evaluation studies that could provide evidence based on experimental studies is not present. The budgets for evaluation are usually far less than 1 per cent of development budgets and the option that they could increase radically does not appear realistic;

2 The Paris Declaration emphasises principles of ownership, harmonisation and alignment, all of

which would make it less likely in the short run that a major share of development spending would occur in the form of experimentally conducted policy initiatives with large sums to be spent simultaneously on experimentally conducted evaluation;

3 There are no signs that the gap between evaluation capacity and evaluation needs, if experimentally conducted studies were to increase rapidly, is being closed. As we saw, it was one evaluation out of 80 that used an experimental design and there were only a few quasi-experimental studies. It is likely that both those who commissioned and those who conducted the other studies were not sufficiently skilled in research design to have initiated experimental or quasi-experimental studies. The capacity – institutional and individual – to conduct experimental and quasi-experimental research is not sufficiently developed.

What is feasible may of course change, and what is not feasible now may be feasible in ten years' time. That depends on whether evidence-based evaluation is desirable and possible.

5.3 Desirable: to some extent 'yes'

This study, as well as many others, has shown that there is a heavy predominance of single narrative analysis in development evaluation. There is not much concern for research design and, as a consequence, we often see an automatic choice of narrative analysis or case studies. A more competitive choice of methods and a higher awareness of methodological choices is likely to provide a higher quality of evaluation. More evidence-based evaluations are thus desirable.

However we would like to make it very clear that the word 'evidence' applies to all forms of research design and that valid and reliable evidence can be produced in different ways – not only through experimental and quasi-experimental designs. Historical research has provided good evidence for the decline and fall of the Roman Empire, palaeontologists have evidence for the extinction of dinosaurs, etc. Evidence for the success or failure of efforts to combat HIV/AIDS in southern Africa is not likely to be generated by the same research methods that generate evidence on the effectiveness of a vaccine or drugs for the infected.

It is the quality of evaluative information that is most important and that all stakeholders need to focus on. The design of the process of inquiry is one aspect of quality, but not the most important one. It is equally important to consider the other choices in the inquiry process; how data are collected and analysed, how samples are made (whether based on experimental or quasi-experimental methods or within case study research), and so on. It is far more important to take a holistic approach to the quality of evidence, irrespective of the design of the process of inquiry. An experimental study can produce poor quality of evidence, as can a case study approach.

Notes

1 Commonly, impact is defined as the long-term, positive or negative, intended or unintended, effects of an intervention (OECD/DAC 2002).
2 A total of 80 evaluations and their terms of reference have been studied and classified according to design, methods, purpose, and subject of analysis. The company, Andante – tools for thinking AB, employed Simon Mikael and Evelina Stadin to organise an information database.
3 We distinguish between *the possible*, which has to do with the practical and theoretical opportunity to do something, for example, if we, as evaluators, could conduct an experimentally designed evaluation responding to the terms of reference, and *the feasible*, which has to with the economical and political constraints to conducting experimentally designed evaluations.

References

DsUD (1991) *Bra Beslut – Om Effektivitet och Utvärdering i Biståndet*, Stockholm: Utrikesdepartementet
OECD/DAC (2002) *Glossary of Key Terms in Evaluation and Results Based Management*, DAC Working Party on Aid Evaluation
Scriven, M. (1991) *The Evaluation Thesaurus*, London: Sage
Vedung, E. (1997) *Public Policy and Program Evaluation*, New Brunswick: Transaction