IDS Working Paper 242

Challenges in evaluating development effectiveness

Howard White

March 2005

INSTITUTE OF DEVELOPMENT STUDIES
Brighton, Sussex BN1 9RE
ENGLAND

Howard White is a Fellow at the Institute of Development Studies (currently on leave of absence at the Operations Evaluation Department, World Bank).

Challenges in evaluating development effectiveness
Howard White
IDS Working Paper 242

First published by the Institute of Development Studies in March 2005

ISBN 1 85864 854 8

A catalogue record for this publication is available from the British Library.

Available from:
Communications Unit
Institute of Development Studies
at the University of Sussex
Brighton BN1 9RE, UK.
Tel: +44 (0)1273 678269
Fax: +44 (0)1273 621202
Email: bookshop@ids.ac.uk
www.ids.ac.uk/ids/bookshop

Printed by XPS Limited, Brighton UK
IDS is a charitable company limited by guarantee and registered in England (No. 877338).

Summary

Evaluation quality is a function of methodological and data inputs. This paper argues that there has been inadequate investment in methodology, often resulting in low quality evaluation outputs. With an increased focus on results, evaluation needs to deliver credible information on the role of development-supported interventions in improving the lives of poor people, so attention to sound methodology matters. This paper explores three areas in which evaluation can be improved. First, reporting agency-wide performance through monitoring systems that satisfy the Triple-A criteria of aggregation, attribution and alignment; which includes procedures for the systematic summary of qualitative data. Second, more attention need to be paid to measuring impact, both through the use of randomisation where possible and appropriate, or through quasi-experimental methods. However, analysis of impact needs to be firmly embedded in a theory-based approach which maps the causal chain from inputs to impacts. Finally, analysis of sustainability needs to move beyond its current crude and cursory treatment to embrace the tools readily available to the discipline.

Contents

Tables

Figures

Acknowledgements

This paper was presented at the OED Conference 'Evaluating Development Effectiveness' in July 2003, and subsequently published as a chapter in G. Keith Pitman, Osvaldo Feinstein and Gregory Ingram (2005), *Evaluating Development Effectiveness,* New Brunswick and London: Transaction Press. Operations Evaluation Department (OED) is gratefully acknowledged for permission to reproduce this paper.

Thanks are due to Osvaldo Feinstein for helpful comments on an earlier version of the paper.

Introduction

The World Bank's mission statement is headed, 'Our dream is a world free of poverty'. The aim of the UK Department for International Development is 'To eliminate poverty in poorer countries' and that of the German Development Agency GTZ is 'To improve the living conditions and perspectives of people in developing and transition countries'. And so on. The Millennium Development Goals (MDGs) are a set of bold promises to poor people around the world. The challenge faced by evaluators is to make credible statements as to whether the activities supported by these agencies make a positive contribution toward their stated goals and so draw lessons to contribute to more effective development.

The nature of this challenge has changed as the development paradigm has evolved, and with it the activities of donor agencies. Stand-alone projects, notably those in infrastructure, can be evaluated using the techniques of cost-benefit analysis that were developed in the 1960s and early 1970s, partly at the World Bank. The shift away from growth as the measure of development was mirrored by a shift in how effectiveness was evaluated. This shift was partly driven by the mistaken view that social sectors were less amenable to economic cost-benefit analysis.[1] It was also driven by a desire to focus directly on non-economic outcomes. The latter was certainly the driving force as the development agenda broadened to encompass rights issues such as gender equality. It was felt that cost-benefit analysis could not possibly capture these things,[2] so that a more qualitative approach was required. By the 1980s, qualitative approaches had come to dominate the evaluation studies conducted for development agencies.[3]

The move toward qualitative approaches was reinforced by an emphasis on process. Process evaluations generally focus on the way in which a project was managed, including aspects such as donor coordination, institutional development, and management systems. These are important issues that may be overlooked in a narrow economic study. And it is usually fair to say that "process projects", whose primary focus is often institutional development, are too far removed from final development outcomes to quantify their impact on the latter.[4]

But the new focus on results in the context of the MDGs demands more than qualitative studies alone can give us. Qualitative methods cannot answer the question, To what extent are agencies' interventions bringing about progress on MDG-related indicators?

---

1 Mistaken since it is precisely the valuation of non-market benefits that provides a motivation for economic rather than financial analysis. Most World Bank appraisal documents for social sector lending state that the rate of return is not applicable. However, Devarajan *et al.* (1996) show that sectoral shifts do not explain the declining use of cost-benefit analysis in the World Bank, which they attribute to the abolition of a central unit responsible for the quality of appraisals and the increased focus on macroeconomic issues in the 1980s.

2 See Kabeer (1992) for an explicit statement of this position.

3 In the words of a former Director General of OED, 'Development interventions today are assessed through a multiplicity of techniques, drawing on many disciplines' (Picciotto 2002: 7).

4 This statement needs to be treated with caution. The ultimate rationale for the assistance is poverty reduction. Where the causal chain is long – such as providing technical assistance to a ministry of finance – there seems little point in attempting to establish any direct link between the activity and poverty reduction. But in other cases, such as supporting the decentralisation of health service delivery – the chain is short enough to make it desirable to be able to say something about how decentralised delivery affects health outcomes.

Quantitative approaches have reestablished themselves at two levels. The first is in measuring agency performance: how well is a particular donor doing? The second is in project-level interventions. Rather belatedly, the development community has been adopting randomisation in programme design to facilitate evaluation. And where this is not possible, recent econometric developments allow the construction of more satisfactory controls than had hitherto been possible. However, the vast bulk of evaluation work is carried out without the use of these techniques.

This paper discusses techniques in three areas of contemporary relevance: measuring agency performance and other studies that aggregate across evaluations; evaluation methods at the project level; and sustainability analysis. In each of these areas it is argued that much evaluation work is carried out well within the "evaluation possibility frontier", which is explored in the following section.

1 The evaluation production function

The argument of the evaluation production function consists of data and techniques, with a positive first differential for each.[5] Figure 1.1 and Figure 1.2 show the combinations of data and technique yielding evaluations of a given quality. In Figure 1.1, a degree of substitutability in production is assumed; for a given budget constraint (the budget for the evaluation), P1, the optimal combination of data and technique can be derived. However, the shortcoming of this analysis is that while data are study-specific, techniques once acquired can be used in subsequent evaluations, which brings down the price of technique (it shifts the bottom end of the budget constraint rightwards along the X-axis). The shift in the budget constraint is greater, the greater the spending on technique in the initial evaluation.

The implications are that optimising quality on an evaluation-by-evaluation basis will result in an overspending on data and an underinvestment in technique. If quality is maximised over more than one evaluation (i.e. a multi-period optimisation), then the resource allocation shifts in favour of techniques, especially in the first of the studies.

It might be thought that substitutability between data and technique is somewhat limited.[6] This view is taken to the extreme of fixed technical coefficients of production in Figure 1.2. Here the optimal combination is always a corner solution, and if the optimal combination is achieved then increases in both data and technique are required to improve evaluation quality. Any resource combination on the vertical segment of the isoquant is inefficient as it entails spending money on data that remain underutilised, given the techniques being applied in the study. And, of course, any additional data collection will add nothing to the study unless there is also some increase in technique. On the other hand, a resource combination

5 One may also consider a third input, theory, which in these figures is subsumed under techniques, to facilitate the graphical presentation.

6 Ravallion (1996) suggests some, but limited, substitutability.

on the horizontal segment of the isoquant has an excess of technique given the available data. I suggest that many evaluation studies are on the vertical segment: they are data-rich but technique-poor. This may seem surprising to economists, since much economic analyses develops very nice techniques but without the data to apply them (that is, they are on the horizontal segment).

Figure 1.1 Evaluation production isoquants (with substitutability in production coefficients)

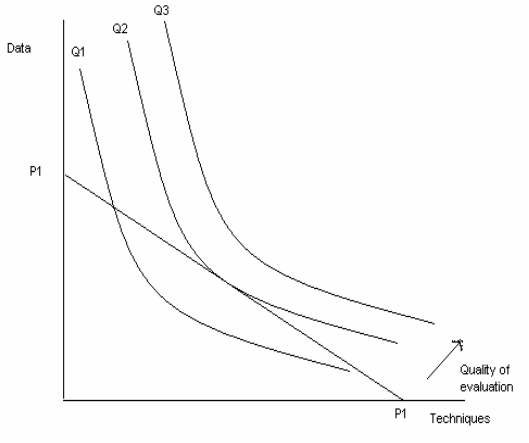
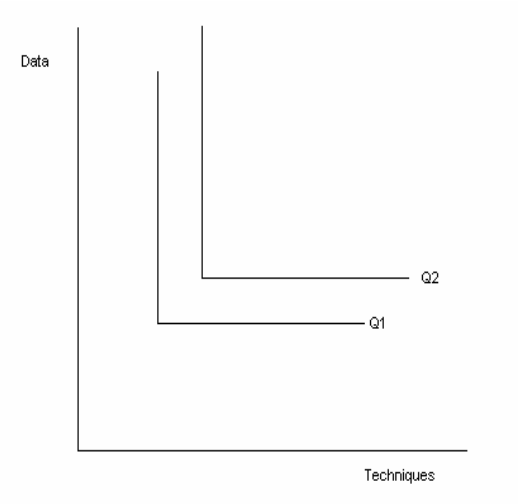


Figure 1.2 Evaluation production isoquants (with fixed coefficients)



Evaluations may also be subject to X-inefficiency. That is, they operate within the evaluation possibility frontier. More explicitly, they are not of the quality that could be achieved given the level of inputs available. Examples of X-inefficiency would be a *mis*use of techniques (rather than their *under*use, in combinations that lie on a non-optimal segment of the evaluation isoquant). For example, a rich data set is

available and used for multivariate analysis, but model misspecification undermines the results. Table 1.1 summarises the possible combinations of misuse and underuse of both data and theory. Ideally a study author can enter 'No' in each cell.

Table 1.1 Possible problems in evaluation production

	Data	Technique/ theory
Misuse		
Underuse		

2 Aggregating results across evaluations

The increased emphasis on results in the context of the Millennium Development Goals has increased the importance of reports that aggregate across evaluations, and the need for these reports to say something about outcomes. Indeed, the MDGs suggest the need for a new sort of report, based on outcomes rather than sectors. The unit of analysis should not necessarily be the sector, but all interventions geared toward achieving a particular MDG.

Inadequate attention has been paid to methodological issues. Below I discuss the lack of explicit attention to the techniques of meta-analysis in studies that involve the aggregation of agency ratings, and in the following subsection, methods used for the collection and analysis of qualitative data, which frequently fall short of what is possible.

2.1 Measuring agency performance

Development agencies are accountable through parliament, their boards, and directly to the taxpayers who finance their activities. In addition to annual reports, which usually focus on a fairly descriptive review of sources and uses of funds by sector and region, several agencies produce more analytical reviews of the development effectiveness of their work. The flagship in this area is the World Bank Operations Evaluation Department's *Annual Review of Development Effectiveness* (*ARDE*), which succeeded the earlier *Evaluation Results* in the mid-1990s. In 1999 the UN Development Program launched its *Results-oriented Annual Report* (*ROAR*), and the UK Department for International Development completed its first *Development Effectiveness Report* (*DER*) for the year 2001. These reports are singled out here as they attempt to summarise agency performance based on aggregations of project-level results.

The World Bank's *Annual Review of Development Effectiveness* uses the ratings of projects that have closed during the year to give an overall portfolio rating of the percentage of projects rated as satisfactory, broken down by country and sector. DFID's *Development Effectiveness Report* uses quantitative data from two reporting mechanisms within DFID: the project completion report and the output-to-purpose review, which is equivalent to the mid-term review in many other agencies. These data are also used to show the

percentage of satisfactory projects. UNDP's *Results-oriented Annual Report* is different, as it draws its data from a new system of reporting geared toward results-based management; no overall performance rating is assigned, but outputs and outcomes are rated against six goals.[7]

Elsewhere I have discussed the triple-A requirements of such reports: attribution, aggregation, and alignment.[8] *Attribution* is the well-known problem of making a link between what the agencies do and the outcomes they hope to influence. *Aggregation* is whether the agency's reporting system produces data that can be meaningfully aggregated across projects. All three agencies measure project performance using scales that facilitate aggregation (for example the percentage of projects falling under a certain classification, such as satisfactory). And *alignment* is whether the data collected at the micro-level tell us anything about performance as measured against the agency's objectives.

All existing systems of agency performance measurement are lacking in some respects against these criteria. DFID's own *Development Effectiveness Report* itself states that there is 'A clear gap between what DFID aspires to achieve, and what it can confidently demonstrate that it has achieved' (p. vi).

The discussion here focuses on just one aspect: the methodology used for aggregations. The reports of each of these agencies aggregate the results of completion reports or other reviews (which here are all called evaluation reports for convenience) to give a summary of portfolio performance. This task is the one faced in conducting meta-evaluation,[9] but the methodology for preparing these reports has not been anchored in the techniques developed for meta-evaluation. Hence it is worth asking if the reports could be improved by a stronger methodology. Here I make some initial observations.

Based on Weiss (1998: 238–9), we can identify six steps to meta-analysis: (1) define the problem; (2) collect the relevant studies to be reviewed; (3) screen the studies for quality; (4) standardise the indicators and code them; (5) produce a synthesis measure; and (6) present results.

The problem addressed in each of these reports is to measure agency performance, while the relevant studies to be used are the agencies' own evaluations. Hence all agencies readily handle the first two steps in a common manner.

The third step, screening out low-quality data, may not seem to be needed since the ratings are assigned by the agency itself. But matters are not so straightforward. *ARDE* uses OED's own ratings based on the department's independent review of the Bank's implementation completion reports. DFID's ratings, by contrast, are those reported in the original project completion report, which is completed by DFID operational staff or the contracted project manager.[10] The guidelines that DFID provides for completing a project completion report are sketchy, and there is no independent check on the ratings

---

7 The six goals are: creation of an enabling environment for sustainable human development, economic and social policies and strategies focused on the reduction of poverty, environmentally sustainable development to reduce human poverty, advancement in the status of women and gender equality, special development situations, and UNDP support to the United Nations.

8 White (2003).

9 The term meta-evaluation refers to the evaluation of evaluations.

10 The team preparing the *Development Effectiveness Report* reviewed 41 project completion reports and 123 output to purpose reviews. They found quality to be very variable, with significant over-grading in project completion reports. However, no corrections were made to the data.

awarded. UNDP's *ROAR* falls between these two extremes. It uses data reported by UNDP operational staff, and it is hoped that staff training will ensure consistency. But those preparing the ROAR screen the data and reject any that fail to meet a minimum standard.[11]

The fourth and fifth steps are the most complicated in meta-analysis, since indicators from different studies may have been measured in different units, using different sample sizes, and have differing levels of statistical significance. But the data used in donor performance monitoring come from common reporting systems, so that no standardisation is required. Moreover, the indicators are not sample estimates, but based on the whole population – for example, in the case of the World Bank, all projects closing in a given fiscal year. So the indicators can be added together, weighting by size if desired, without the need for statistical adjustments. This would not be the case if it were decided to use a sample of project ratings, rather than those from the whole portfolio.

Step five, the production of a synthesis measure, raises the issue of whether like is being added to like, or if apples are being added to oranges. Adding apples and oranges is to some extent inevitable given the diversity of the portfolio. "Satisfactory" cannot mean exactly the same in a road-building project as it does in a financial sector adjustment credit. In this case comparability means that we are satisfied that common criteria have been applied in awarding the ratings, something that is ensured by well-established guidelines and an independent check. Beyond that, it *is* okay to add apples and oranges if we want to know how many pieces of fruit we have.

For the purpose of quantitative portfolio reviews, then, this brief discussion identifies some, though rather limited, gains that could be achieved from the formal application of meta-analysis techniques.

However, none of the reports is restricted to the synthesis of quantitative ratings data. OED, for example, produces country assistance evaluations and sector studies, and each year OED's *Annual Review of Development Effectiveness* draws lessons from these. Similarly, DFID's *Development Effectiveness Report* reviews progress against the goals set out in DFID's institutional strategy papers.[12] Meta-analysis also requires the aggregation of qualitative data, which the following discussion suggests has proved more problematic.

2.2 Country and sector studies

Each year the World Bank's Operations Evaluation Department conducts a number of country assistance evaluations and sector studies. Two issues relating to such studies are addressed here: (1) the role of case studies, and (2) summarising qualitative data.

---

11 United Nations Development Program (2000: 63).

12 The institutional strategy papers lay out DFID's vision of the appropriate role of the international agencies with which DFID works, and how DFID may best work with that agency. For each institutional strategy paper there will be an institutional performance review that can be used in an evaluation synthesis, though at the time of preparing the first development effectiveness report only one institutional performance review was available.

2.2.1 The role of case studies

Sector and country studies combine a number of approaches. The portfolio review analyses lending trends in the Bank's portfolio for that sector or country, summarises ratings-based performance measures, and may analyse aspects of design. The background work for an OED sector study will typically include four to six country case studies, as do many large evaluation studies by other donors (e.g. DFID's poverty evaluation and the Swedish International Development Agency's global evaluation of programme aid). The choice of countries is frequently contentious, with critics arguing that nothing can be learned from looking at just four countries when the portfolio covers 60 or more.

OED and other evaluation departments use two well-established defenses of the combined methodology they use for sector and country studies. The first is the familiar debate on a case study approach, which abandons representativeness (breadth) for depth. The depth of engagement in a case study allows an exploration of issues that cannot possibly be uncovered in a desk-based portfolio review (which may be seen as synonymous with a cross-country approach).[13] The second defense is that the study is not based on the case studies alone but uses a mix of methods, combining the data from a number of instruments and approaches.[14]

OED uses portfolio reviews, for example, in both sector studies and country assistance evaluations. At the very least these reviews describe trends in the size and composition of the portfolio (for example by region or subsector) and an analysis based on OED ratings. But they may be more comprehensive, studying the evolution of project objectives and design over time.[15] In the portfolio review in OED's study of social funds,[16] design features such as the use of outreach, targeting mechanisms, community contribution, and maintenance arrangements were codified based on a desk review supplemented with task manager interviews. The theory-based approach used in this evaluation suggested the importance of some features for successful operations, such as comprehensive outreach and community-level arrangements for maintenance. Four country case studies were undertaken to provide material with which to test the importance of these factors, and their results helped identify the most critical design features. The portfolio review showed how social funds had evolved, acquiring better design features over time.

2.2.2 Summarising qualitative data

As is well known, mean-based quantitative statistics can give a misleading summary. For example, in a rather bizarre week in Washington in April this year it snowed at the beginning of the week, with the temperature 15 degrees below the usual average for April. But by the end of the week it had reached 15 degrees above average. So it can safely be said that the average temperature that week was equal to the usual April average! This statement is true but misleading; it is not a good summary of the data.

---

13 See Casley and Lury (1987) for a fuller discussion of the role of case studies.

14 See Rao and Woolcock (2003) for a more general discussion of the iterative approach to combining quantitative and qualitative methods in evaluation.

15 As is the case, for example, in OED's health sector review, World Bank Operations Evaluation Department (1999).

16 World Bank Operations Evaluation Department (2002).

Regression coefficients are also a mean-based statistic, and failure to inspect the data can give misleading results if influential points "distort" the slope of the line or if an incorrect functional form is used.[17]

Equally, techniques should be used to ensure that qualitative data are summarised in a way that reveals, rather than distorts, the patterns in the data. The failure to use a systematic approach may lead to "cherry picking". For example, suppose six country case studies are conducted as part of an evaluation study. One of the study questions is whether decentralisation increases the feeling of autonomy of local government officials.[18] The study design has not included structured questionnaires but unstructured interviews with officials in five districts in each country. It may well be possible to write that: 'Decentralisation increases the feeling of autonomy of local government officials; for example, one local official in Malawi commented that "The changes have made a real difference, I really feel in control of what I do now"'. But it may be that no one else in Malawi made a similar comment, and nor did any official in any of the other five countries. Such a misrepresentation of the data is an example of data mining (searching the data until you find the evidence you are looking for) – though we should recognise that this is hardly something of which quantitative analysts are innocent.[19]

Well-established methods are available for the systematic analysis of qualitative data; they include content analysis and computer programmes (e.g. QSRNudist) for conducting such analysis.

Yet this is an area where many evaluations either fall within the evaluation possibility frontier or devote inadequate resources to technique. It is rare to find evidence that qualitative data have been collected and analysed in a systematic way.[20] For example, in a multi-county study, it is not sufficient to have common terms of reference. More detailed guidance is required on the range of data to be collected and the sources to be used. And these data need be presented in such a way as to allow a systematic summary.

Where there are a small number of country case studies I believe that a tabular summary can suffice. Returning to the decentralisation example, suppose we are interested in three questions: (1) does decentralisation increase the feeling of autonomy of local government officials? (2) does decentralisation increase the perceived accountability of local government among local communities? and (3) does decentralisation result in an increase in spending on community-level facilities? In all six countries, data have been collected on these issues through, respectively, (1) interviews with local officials, (2) focus group and key informant interviews in communities and, (3) budget analysis from five districts in each country, supplemented by the key informant interviews at community level. A table can then be constructed with the following structure:

---

17 For example, fitting a straight line to a quadratic relationship can yield an insignificant coefficient even if the true relationship is very strong.

18 This is a hypothetical example; it does not refer to an actual study.

19 For an elaboration of this point see White (2001). Data mining should be distinguished from the perfectly legitimate enterprise of data analysis, a data-driven approach that allows the data to tell their own story rather than being forced into a preconceived view of what they should say (i.e. a pre-specified model). See Mukherjee *et al.* (1998).

20 I do not pursue here the very scant use made in some official agencies of participatory evaluation, a particular approach to qualitative data that has wide support amongst non-governmental agencies.

Table 2.1 Schema for summarising qualitative data

	Country a	Country b	Country c	Country d	Country e	Country f
Question 1						
Question 2						
Question 3						

The cell summarises the evidence from that country for that question. Given a systematic approach to data collection, no cell should be left blank. Reading across the rows gives an overall impression of the weight of the evidence. This evidence should be reported in a way to show that all the data are being summarised: 'In five of the six countries no evidence was found that decentralisation had increased feelings of autonomy of local officials, and in the sixth country (Malawi) only one official mentioned a positive effect'. In the executive summary it would suffice to say that 'The evidence collected by the study did not support the view that decentralisation increased the feeling of autonomy amongst local officials'. If the exception is to be mentioned at all, it should not be in a way that gives it undue emphasis.[21] The systematic recording of data can also help avoid misreporting findings.

The misrepresentation of data by cherry picking is sometimes formalised in the best practice approach, which focuses the analysis on an acknowledged outlier. Studies may focus on desirable processes and impacts. But they frequently do not document how exceptional is this best practice or examine the conditions that explain why best practice has been achieved in one case but not others. At its worst, focusing only on best practice can result in an entirely unrealistic approach to project and programme design. There is of course some merit in looking at best practice. But such analysis must involve looking at the conditions required for best practice and their replicability. There is also a case for looking at worst practice – don't we want to learn from our mistakes? To put it another way, we want to know what works and what doesn't. In order to learn that, we must look at cases that have not worked.

3 Micro-level evidence of development effectiveness

The term impact evaluation has been used with several different meanings. The four most common, which are not mutually exclusive, are:

- Rigorous analysis of the counterfactual;
- A focus on outcomes;
- Evaluation carried out some years after the intervention has ended;
- Country or sector-wide studies.

[21] For example, 'In Malawi evidence was found that decentralisation has increased the feelings of autonomy of local officials, although the evidence was less strong in the other countries'. The quantitative equivalent of this Malawi example is either basing arguments on statistically insignificant coefficients or allowing a single influential point to drive the results.

A review of OED's 108 impact evaluation reports listed in the World Bank's document database[22] shows that all these different meanings of impact have been used.[23]

In the results-based climate of today, impact evaluations focus on outcomes. In any study they conduct, evaluators should be concerned with the requirement to take account of the counterfactual. Within this context an increasing number of studies, many from the World Bank's research department, have presented a counterfactual analysis of outcomes. These studies typically rely on advanced econometric techniques, which are being promoted in some quarters as the model to adopt for all evaluations, and certainly for evaluations with an impact focus. But such techniques may not always be appropriate, so there is still a need to develop rigorous alternative approaches.

3.1 Technical innovations for measuring impact: randomisation and propensity score matching

At the micro-level the problem of attribution is embodied in the construction of a counterfactual: what would outcomes have been in the absence of the intervention? Common ways of addressing this issue are:

- Before versus after comparison of outcome indicators for those benefiting from the intervention (commonly called the treatment group). Since this method does not control for other factors affecting outcomes, it is not usually regarded as giving credible results.
- Comparison of outcomes with those in a control group. This approach requires a control group whose outcomes were similar to those of the treatment group prior to the intervention and which has subsequently been subject to the same shocks and trends as the treatment group.
- The double difference method, which combines the previous two, by comparing the change in the outcome in the treatment group with the change in the control group. This method is preferred because it eliminates constant unobservable differences between the treatment and control groups.

The problem is of course to identify a credible control group. One way of doing this is to allocate programme resources in a random manner. In that way, programme beneficiaries are a random sample of the population as a whole, and their outcomes can be compared with those of another randomly drawn sample of non-beneficiaries (the control group).[24] Randomised design has been widely used in medicine

---

22 Gupta Kapoor (2002).

23 The rather diffuse meaning of impact in these studies was one source of criticism from the Meltzer Commission: 'The banks seldom return to inspect project success or assess sustainability of results. The World Bank reviews only 5 per cent of its programs three to ten years after final disbursement. These impact evaluations focus on such important, but poorly defined and subjective, measures as improvements in the environment, the role of women, the interaction of societal institutions, income distribution and general welfare. It is difficult to relate Bank activities to these social indicators. Thirty per cent of the investigators found that lack of monitoring of project results precluded valid judgments. Though the agencies devote significant resources to monitoring the procurement of inputs, they do little to measure the effectiveness of outputs over time.' International Financial Institution Advisory Committee (2000).

24 The unit of observation need not be the individual. It could for example be a school, a district or a farmers' group.

and is increasingly used in social interventions in developed countries. Its use in developing countries is more recent, as reviewed in the papers by Rawlings (2005) and Duflo and Kremer (2005).

Where randomisation is not possible (or may have been possible but the intervention was not designed that way)[25] then the comparatively recent technique of propensity score matching allows the construction of good controls.[26] The objective here is to match each individual in the treatment group with a person in the control group, producing pairs of individuals or households who are as similar to one another as possible. The data on which we wish to match the individuals or households are a set of observable characteristics that are unaffected by participation in the programme. Thus baseline data are a first requirement. But even with just, say, five characteristics, matching becomes difficult and ultimately arbitrary when several dimensions are being used to make the match. As Rosenbaum and Rubin (1983) showed, we can match by the propensity to participate in the programme, where this propensity is the predicted probability of programme participation based on the observed characteristics.[27] Each participant can then be matched with the non-participant who has the nearest "score" and the difference in outcomes in the pair calculated. In practice, results are better when matching with a larger number of members from the control group, such as the nearest five,[28] or when matching each participant with all members of the control group, calculating the mean outcome for the control group using weights that vary according to the similarity in score.

Taken together, these techniques provide a firm basis for saying something about impact. The problem is that they are applicable only to only those types of intervention that can be seen as delivering a treatment to a clearly defined group. The medical analogy is appropriate. Where an intervention is discrete and homogenous, like taking a tablet, then these techniques can be used. But where this is not the case, alternative approaches are needed.

3.2 Alternative approaches to impact evaluation

These techniques are not suitable for a wide range of activities that development agencies support, including the policy reforms supported by adjustment lending and the programme aid of bilateral donors, or other channels through which policy advice is given, such as economic and sector work.[29] Nor do they apply to technical assistance that is provided to support institutional development rather than to implement specific activities. Nor can they be used to evaluate institutional development activities more

---

25 Randomisation is in principle possible for all interventions for which the universe of potential beneficiaries exceeds the number of possible beneficiaries allowed by the programme budget (or the possible number in the first phase). There are cases in which all possible beneficiaries are intended to be reached – for example the HIPC initiative, recently evaluated by OED (World Bank 2003).

26 See Ravallion (2002) for an accessible introduction to the technique.

27 That is, we estimate a logit model for participation, in which the dependent variable is 1 for participants and 0 for non-participants and the regressors are the observed characteristics.

28 In this case, one calculates the mean outcome of the nearest five and compares this mean with the outcome for the participant.

29 There may be rare cases of sectoral policies that can be implemented in discrete geographical units on a random basis, but these would be very much the exception. But if policy change is costless, the moral defense for withholding beneficial treatment from the control collapses.

generally where these are carried out either at the central level or in a very small number of districts. For such cases, alternative approaches are needed.

Furthermore, match-based comparisons assume that all beneficiaries receive the same treatment. This is not the case in most interventions, and it is helpful to know which parts of project design work and which do not. An example of the problem of heterogeneity is provided in a recent analysis of social funds. This study used household-level data from six countries to conduct regressions of welfare outcomes. [30] A social fund dummy variable was assigned a value of one if the household lived in a community with a social fund intervention in the appropriate sector and a value of zero otherwise. There are three problems in this approach. First, we do not know if "non-social fund communities" have no facility or a facility that has not received social fund support.[31] Hence we do not know what is the counterfactual. Second, the nature of social fund support can vary in scale and type between facilities. Classroom rehabilitation will have less impact on enrolments than expanding the size of the school by building new classroom blocks, but all are subsumed under the project dummy. Finally, the use of the dummy variable does not allow us to unpack the causal chain to understand why effects are, or are not, being found. Several of the regressions showed an insignificant "social fund effect" and even in a few cases a perverse one. For example, in Zambia health facilities supported by the social fund exacerbate wasting in the recipient community.[32] It is more useful to "open the black box" of the project contained in the use of the project dummy.

The methodology in OED's current programme of impact evaluations is based on modeling the determinants of those outcomes given by the objectives, and then linking the outputs of Bank interventions to those determinants.[33] Applying this logic to the study of social funds just mentioned, it would have made sense to model the determinants of the welfare outcomes, where some of these determinants (such as access to and quality of services) could be linked to the social fund intervention. An advantage of this approach is to merge a process-oriented approach with impact analysis. The causal chain included in the logical framework encompasses process issues, such as the dynamic behind policy changes, some of which may be addressed using qualitative techniques.[34]

---

30 Some of the results from these country studies are presented in *World Bank Economic Review* (August 2000).

31 In some studies the selected communities were those in the pipeline. The rationale for using pipeline communities is that their acceptance for a project should mean they are similar to already accepted communities, but this fact does not solve the problems mentioned here.

32 These results are summarised in World Bank (2002), Annex E.

33 Earlier OED impact evaluations also adopted non-experimental approaches (the most recent examples are World Bank 1998 and 2000) and see Gupta Kapoor (2002) for a review.

34 See World Bank (forthcoming) for an illustration.

4 Sustainability

Sustainability is often equated with environmental concerns. But in evaluation, environmental sustainability is just one aspect of sustainability. For example, the OECD Development Assistance Committee glossary states that 'Sustainability is concerned with measuring whether the benefits of an activity are likely to continue after donor funding has been withdrawn. Projects need to be environmentally as well as financially sustainable.'[35] OED defines sustainability as 'The resilience to risk of net benefits flows over time', elaborating the definition with the following questions: 'At the time of evaluation, what is the resilience to risks of future net benefits flows? How sensitive is the project to changes in the operating environment? Will the project continue to produce net benefits, as long as intended, or even longer? How well will the project weather shocks and changing circumstances?'[36]

However, beyond definitions there is little agreement. A recent review stated that 'Much remains to be done in terms of [sustainability's] evaluation as an objective'.[37] The OED definition makes clear the strong link between sustainability analysis in evaluation and risk analysis at appraisal. In principle it should be possible to use the same techniques for both. However, when we turn to the World Bank's practice in appraising risk, we find it falls far short of best practice.

The World Bank's Operational Policy 10.04 (Economic Evaluation of Investment Operations) deals with both sustainability[38] and risk.[39] The approach it advocates is traditional deterministic sensitivity analysis, that is, varying key assumptions and noting the changes in the return to the project.[40] But, as argued by Belli *et al.* (2000), this approach is of limited use, since it does not tell us how likely or how large is the variation from our assumed value. In addition, varying one assumption at a time understates the risk if there is a positive correlation between changes in the different variables. The preferred approach is to

35 Development Assistance Committee Working Party on Aid Evaluation (2002).

36 www.worldbank.org/oed/eta-approach.html

37 CIDA (2002: 1).

38 'To obtain a reasonable assurance that the project's benefits will materialize as expected and will be sustained throughout the life of the project, the Bank assesses the robustness of the project with respect to economic, financial, institutional, and environmental risks. Bank staff check, among other things, (a) whether the legal and institutional framework either is in place or will be developed during implementation to ensure that the project functions as designed, and (b) whether critical private and institutional stakeholders have or will have the incentives to implement the project successfully. Assessing sustainability includes evaluating the project's financial impact on the implementing/sponsoring institution and estimating the direct effect on public finances of the project's capital outlays and recurrent costs.'

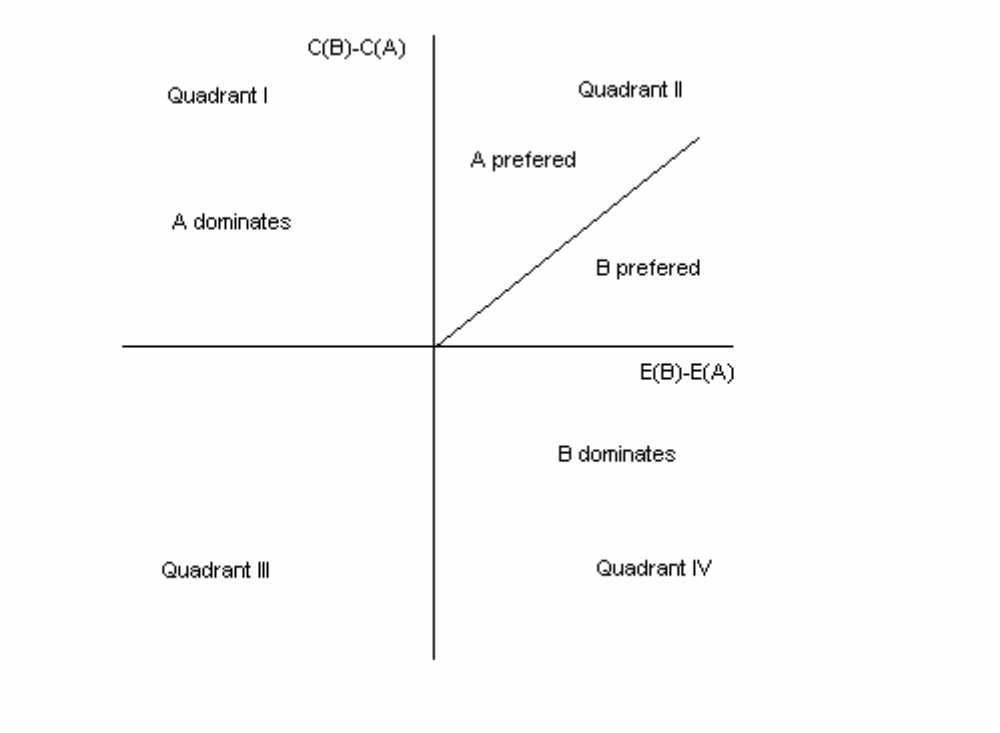
39 The economic analysis of projects is necessarily based on uncertain future events and inexact data and, therefore, inevitably involves probability judgments. Accordingly, the Bank's economic evaluation considers the sources, magnitude, and effects of the risks associated with the project by taking into account the possible range in the values of the basic variables and assessing the robustness of the project's outcome with respect to changes in these values. The analysis estimates the switching values of key variables (i.e. the value that each variable must assume to reduce the net present value of the project to zero) and the sensitivity of the project's net present value to changes in those variables (e.g. delays in implementation, cost overruns, and other variables that can be controlled to some extent). The main purpose of this analysis is to identify the scope for improving project design, increase the project's expected value, and reduce the risk of failure.

40 Deterministic sensitivity analysis can be either univariate – varying one assumption at a time – or multivariate – varying several assumptions at once (in different scenarios).

conduct Monte Carlo simulations, which requires specifying the distribution of all-important variables and the correlation between them.[41] In each simulation the internal rate of return is calculated so that the frequency distribution of the return can be plotted.

Recent papers in health economics have adopted a Bayesian approach to sensitivity analysis.[42] Health evaluations use the incremental cost-effectiveness ratio (ICER), which may be seen as equivalent to the familiar cost-benefit ratio in economics.[43] The ICER compares a new treatment (B) with an existing one (A), or compares pairs of possible treatments. The numerator is the difference in the cost of the treatments, which includes cost savings from health improvements from the new technique. The denominator is the change in some measure of health outcomes, such as quality-adjusted life-years (QALYs). The result is a four-quadrant diagram (Figure 4.1). In quadrant I the old treatment dominates, as it was cheaper and more effective, whereas in the fourth quadrant the new treatment dominates. In quadrant II the new treatment is more effective but also more expensive. It will be preferred if the ratio exceeds some norm on acceptable cost per QALY saved. In the third quadrant the new treatment is cheaper but less effective, which gives ambiguous feelings about what to do.

Figure 4.1 Analysis of the incremental cost-effectiveness ratio



41 A Monte Carlo simulation simply means repeating an event with a random outcome many times. The simulations themselves are relatively straightforward on a spreadsheet. The more difficult task is to specify the underlying distribution and the correlation between variables. See the discussion in Belli *et al.* (2000) Chapter 11 on how to do this.

42 See for example Briggs (1999).

43 If a "no project" counterfactual is used, as is often the case in cost-benefit analysis, then the values for the comparison in the cost-benefit ratio are zero.

Deterministic sensitivity analysis can be applied to calculating the incremental cost-effectiveness ratio, as can a stochastic approach that maps its distribution. The problem with the latter is that the ICER is non-bounded for small changes in effects and that equal values of the ICER from the second and third quadrants actually refer to rather different things. The Bayesian approach thus takes prior assumptions about model values to calculate the posterior distribution of the ICER.[44] Specifically, it is possible to calculate the probability that B will be preferred over A (probability in quadrant IV plus probability in lower part of quadrant II). Other studies calculate the probability of the ICER being in each quadrant separately and then map the distribution in that quadrant.

All this is rather far from current practice at the World Bank. World Bank appraisal documents at best follow the Bank's guidelines of conducting deterministic sensitivity analysis. None calculates a risk-based expected return.[45]

Hence there is an established method of tackling sustainability in project appraisal that could equally well be used at the evaluation stage in re-estimating the IRR. However, there are three reasons for not following this route to sustainability analysis.[46] First its relative technical sophistication may mean that it is only used on a small selection of projects, which would be disadvantaged by this more rigorous analysis compared with other less closely scrutinised projects. Second and more importantly, the method can produce impressive-looking output that may distract users of evaluation reports from the critical step of identifying key assumptions and their attendant probability distributions. Third, such analysis is most amenable to varying assumptions on clearly defined variables, such as price and yields, and it can readily incorporate delays. It is less easy to see how to incorporate the types of risks often identified in project documents, such as "government will not maintain its commitment". As an intermediate step, it is proposed that analysis of sustainability would be best served by more serious attention to risks and their likelihood. When this becomes established practice then it may be time to move to complete the formalisation of the approach by linking these risks to variables used in the calculation of the rate of return and producing distributions of those returns.

How should key risks be identified? Belli *et al.* (2000) suggest that sensitivity analysis can be used to identify which variables matter most. But this approach will only work if a spreadsheet project analysis has been set up. And it will not capture such factors as "lack of government commitment". More appropriate is a theory-based approach similar to a log-frame analysis of project design, which should already

---

44 Model specification is likely to be the most important source of uncertainty.

45 A free text search of the term "expected net present value" in the World Bank's document database, Imagebank, returns only one project appraisal document. In this case the term expected was being used in the lay sense, that is, "given our assumptions we expect the IRR to be this", rather than its more precise statistical meaning.

46 These suggestions apply equally to appraisal and evaluation.

Table 4.1 Possible factors in sustainability analysis

Aspect	Factor	Data
Financial		
Government	Strength of tax base	Historical analysis of revenue variation and factors behind it (e.g. commodity prices).
	Ability to finance	Shortfalls between actual and budgeted expenditure.
	Willingness to finance	Likely government priorities. If government may change, what are priorities of opposition parties?
Communities	Ability to pay	Forecasts of charge as a percent of average income (requires analysis of livelihood of community).
	Willingness to pay	Possible changes in pattern of demand, perhaps because of competing suppliers.
		Will the project continue to produce benefits of sufficient value? (e.g. prices fall so output not worth as much, so producers won't pay for technical services).
Donors	Continued donor finance	Is it planned?
Institutional		
Government	Capacity	Are the skills required for project implementation or supervision present and will they stay there?
	Existence	If the project is under a special unit, will it continue to exist or are arrangements in place for the takeover of functions by exiting government units?
NGO	Organisational viability	Will the organisation continue to exist without external project finance?
	Capacity	Are the skills required for project implementation present and will they stay there?
Community	Organisation	Are there command-based organisations to be responsible for community level responsibilities? Can these responsibilities be enforced?
	Capacity	Do community members have the required skills, or are mechanisms in place so that they acquire them?
Technical		
Design	Soundness	Are physical structures sound or do they have structural problems?
Operations and maintenance	Ability	Do those responsible have the required skill for operation and maintenance?
Environmental		Is project depleting stocks of a non-renewable source? Are there adverse environmental consequences of project technology that will undermine its effects?

incorporate risk. Theory-based evaluation (TBE) seeks to uncover the key assumptions that underlie project design. It asks, 'The goal is to improve indicator Y, and the project is delivering input X. What is the causal chain, or set of assumptions, by which the project designers believe that input X will affect outcome Y?'. A theory-based evaluation tests these links.

TBE can be adapted to the analysis of sustainability. The theory lays out a set of conditions under which X will lead to Y. If Y cannot be observed – as it cannot in the case of future benefits – then the evaluation can test whether or not the conditions for X to lead to Y are in place. Whereas TBE per se

seeks to test these conditions, the analysis of sustainability uses reference to other literature to validate the theory linking X to Y. This approach was adopted in OED's analysis of the sustainability of sub-projects supported by social funds.[47] Based on a review of the literature on project sustainability, conditions were derived that should be met if sub-projects are to be sustainable, and the importance of these conditions was validated through the four country case studies. The portfolio review recorded which projects had the required design features for sustainability and which did not. The study found that social funds were learning the lessons of experience and beginning to adopt the required design features for sub-project sustainability more widely.

Theory-based evaluation enables identification of factors critical for sustainability, but it does not tell us about the probability of different events. It would seem a minimum requirement that identified risks should be labeled as "very likely" to "not at all likely", with some common understanding of the range of probabilities referred to by each label. Where the risks are of a vague sort ("lack of government commitment") the evaluator should be required to spell out in more detail the implications for the project. Table 4.1 provides an example of some reasonably generic factors that might be considered in a sustainability analysis.

5 Conclusion

Evaluation has a crucial role to play in today's results-based culture. Studies must be able to credibly establish the beneficial impact on the poor of official interventions. And they must be able to draw out relevant and applicable lessons for policymakers. Evaluations frequently fall short of their potential, usually because their authors give inadequate attention to application of the most appropriate techniques.[48]

This paper has discussed the following cases of the need for more attention to technique:

- formal application of meta-analysis in studies that aggregate performance (agency-wide performance, and country and sector studies);
- weak analysis of qualitative data, including the prevalence of data mining (cherry picking), sometimes formalised in the best-practice approach;
- paying greater attention to establishing the control in evaluation design, either through randomisation where possible or through propensity score matching – both of which imply taking a prospective approach, which is not commonly applied;
- seeking ways to establish impact that "open the black box" and so provide lessons about what works and what doesn't; and
- application of risk analysis to discussions of sustainability.

---

47 World Bank (2002).

48 Not discussed here is the preference for "quick and dirty" studies as many agencies shy away from the higher cost of "longer but clean" studies. However, any cost-benefit calculation will favor the latter.

This is a demanding agenda. But results-orientation is demanding, and the challenge of eliminating world poverty more so. Evaluation units and evaluators in the development community need to ask themselves if they are doing their best to help meet this challenge.

References

Belli, P., Anderson, J., Dixon, J and Jee-Peng, T., 2000, *Economic Analysis of Investment Operations: Analytical Tools and Practical Applications*, Washington, D.C.: World Bank Institute, World Bank

Briggs, A., 1999, 'A Bayesian approach to stochastic cost-effectiveness analysis', *Health Economics Letters,* Vol 8: 257–61

Canadian International Development Agency (CIDA), 2002, 'Assessing sustainability', *What We're Learning* 2, Ottawa: Canadian International Development Agency

Casley, D.J. and Lury, D.A., 1987, *Data Collection in Developing Countries,* Second Edition, Oxford: Oxford University Press

Devarajan, S., Squire, L. and Suthiwart-Narueput, S., 1996, 'Project Appraisal at the World Bank', in C. Kirkpatrick and J. Weiss, *Cost-Benefit Analysis and Project Appraisal in Developing Countries,* Cheltenham: Edward Elgar

Development Assistance Committee Working Party on Aid Evaluation, 2002, *Glossary of Key Terms in Evaluation and Results-based Management*, Paris: Development Assistance Committee, Organisation for Economic Cooperation and Development

Duflo, E. and Kremer, M., 2005, 'Use of Randomization Evaluation of Development Effectiveness', in G.K. Pitman *et al., Evaluating Development Effectiveness,* New Brunswick and London: Transaction Press

Gittinger, J., 1985, *Economic Analysis of Agricultural Projects,* Washington, D.C.: Economic Development Institute, World Bank

Gupta Kapoor, A., 2002, 'Review of the Impact Evaluation Methodologies used by the Operations Evaluation Department over the past 25 years', *OED Working Paper*, Washington, D.C.: World Bank

International Financial Institution Advisory Commission (IFIAC), 2000, *Final Report of the International Financial Institution Advisory Commission to the US Congress and Department of the Treasury,* Washington, D.C.: IFIAC

Kabeer, N., 1992, 'Evaluating cost-benefit analysis as a tool for gender planning', *Development and Change* 23: 115

Little, I.M.D. and Mirrlees, J.A., 1990, 'Project appraisal and planning twenty years on', paper prepared for World Bank Conference on Development Economics, April

—— 1974, *Project Appraisal and Planning for Developing Countries,* London: Heinemann Educational

Mukherjee, C., Wuyts, M. and White, H., 1998, *Econometrics and Data Analysis for Developing Countries*, London: Routledge

Picciotto, R., 2002, 'Development cooperation and performance evaluation: the Monterrey challenge', *OED Working Paper*, Washington, D.C.: World Bank Operations Evaluation Department

Rao, V. and Woolcock, M., 2003, 'Integrating Qualitative and Quantitative Approaches in Program Evaluation', in F.J. Bourguignon and L. Pereira da Silva (eds), *The Impact of Economic Policies on Poverty and Income Distribution,* New York: Oxford University Press: 165–90

Ravallion, M., 2002, 'The mystery of the vanishing benefits: an introduction to impact evaluation', *World Bank Economic Review,* Vol 15 No 115

—— 1996, 'How well can method substitute for data? Five experiments in poverty analysis', *World Bank Research Observer,* Vol 11 No 2: 199–221

Rawlings, L., 2005, 'Operational Reflections on Evaluating Development Programs', in G.K. Pitman *et al., Evaluating Development Effectiveness,* New Brunswick and London: Transaction Press

Rosenbaum, P. and Rubin, D., 1983, 'The central role of propensity score matching in observational studies for causal effects', *Biometrika* 70: 41–55

Squire, L. and van der Tak, H., 1975, *Economic Analysis of Projects,* Washington, D.C.: World Bank

United Nations Industrial Development Organization (UNIDO), 1972, *Guidelines for Project Evaluation,* New York: United Nations Industrial Development Organization

UNDP, 2000, *Results-Oriented Annual Report,* New York: United Nations Development Programme

Weiss, C., 1998, *Evaluation,* New York: Prentice Hall

White, H., 2003, 'Using the MDGs to Measuring Donor Agency Performance', in R. Black and H. White, *Targeting Development: Critical Perspectives on the Millennium Development Goals,* London: Routledge

—— 2001, 'Combining quantitative and qualitative techniques in poverty analysis', *World Development,* Vol 30 No 3: 511–22

World Bank Operations Evaluation Department, (forthcoming), *Adjusting Education: An Impact Evaluation of World Bank Support to Basic Education in Ghana,* Washington, D.C.: World Bank Operations Evaluation Department

—— 2003, *Debt Relief for the Poorest: An OED Review of the Highly Indebted Poor Countries Initiative*, Washington, D.C.: World Bank Operations Evaluation Department

—— 2002, *Social Funds: Assessing Effectiveness,* Washington, D.C.: World Bank Operations Evaluation Department

—— 1999, *Investing in Health: Development Effectiveness in Health, Nutrition, and Population,* Washington, D.C.: World Bank Operations Evaluation Department

—— 1998, *India: The Dairy Revolution,* Washington, D.C.: World Bank Operations Evaluation Department