Using Strong Evaluation Designs in Developing Countries: Experience and Challenges

Michael Bamberger and Howard White
Independent Consultant and Independent Evaluation Group, The World Bank

In his article in the November 2006 *JMDE* Thomas Cook reviewed the current debate on the role of randomized control trials (RCT). While his article focused on educational research, the issues raised, and the responses by Michael Scriven (2006) and Jane Davidson (2006), highlighted the broader issues currently being debated on the need for more 'rigorous' program evaluation in all sectors. The purpose of this article is to extend the discussion to the field of international development evaluation, reviewing the different approaches which can be adopted to rigorous evaluation methodology and their applicability in a development setting.

In the international development field there is a growing debate on the need for more rigorous evaluation design, and specifically the use of randomized control trials. The majority of evaluations carried out by official development agencies have largely been process evaluations. In addition, from the early eighties there was an increased use of participatory evaluation approaches. Such approaches were a welcome change from a tradition which had not seen it necessary to seek the opinion of the intended beneficiaries of aid-financed interventions, but they did not produce data amenable to quantitative analysis of impact. The rise of results-based approaches in government agencies in developed countries, and the associated focus on the Millennium Development Goals as a measure of development progress, has resulted in greater calls to be able to demonstrate aid impact, and to know what works (see White, 2004, for a discussion of the results agenda in development agencies).

Several international conferences have stressed the need for greater accountability in the use of aid and greater rigor in the assessment of development outcomes. In particular, the 2002 Monterrey Conference on Financing for Development heightened interest in the use of results-based management in development agencies and the 2005 Paris Accords encouraged multi-donor cooperation in the promotion of, among other things, impact evaluations. The Poverty Action Lab (PAL) at MIT has for a number of years been promoting the use of randomized designs and also offers training programs for developing countries on the use of these designs. The Center for Global Development (CGD) has become a strong advocate of more rigorous evaluation designs, notably in their publication "When Will We Ever Learn" (CGD, 2006). Recently CGD issued a "Call to Action", calling for the creation of an independent evaluation agency to ensure more independence and rigor in the evaluation of development programs. The term "Gold Standard" has recently been introduced into evaluation discourse to refer to RCTs as being the impact evaluation methodology to which development agencies should aspire, though this privileged position is disputed by others.

The purpose of this article is to seek common ground on ways to strengthen the

methodological rigor and quality of development impact evaluations, while at the same time adapting the methodology to the technical, administrative, political and socio-cultural contexts within which these evaluations are developed, implemented and used.

The Strengths and Limitations of RCTs in the International Development Context

Cook argues that evaluation has to deal with different kinds of ideas and issues and causal questions are not always important. This is indeed the case, as shown by the heavy reliance of development agencies on process evaluations for lesson learning and accountability (with respect to the use of inputs and possibly assessing actual outputs against target values). However, for answering questions about causality randomized designs can be well warranted theoretically and empirically and they are widely perceived to have at least a marginal advantage over other bias-free methods such as regression discontinuity. Experiments also enjoy credibility in some policy debates—though this is not yet the case in most developing countries or even in all international development agencies. Consequently, having rigorous evidence on impact can increase the likelihood of information being used in policy debates.

Cook argued that randomized designs have a significant methodological advantage over quasi-experimental designs, none of which can adequately control for sample selection bias. Cook recognizes that valid knowledge has often come from non-experiments and there are many situations in which randomized designs are not possible. Despite their advantages, relatively limited use has been made of randomized designs in social (policy) research. He concludes that multiple methods are required in policy research, that generating better causal knowledge has a large role in policy research and that there is a special need for experiments today.

Many of these arguments are equally applicable to the international development field. International development projects typically use one of two procedures for participant selection: self-selection (people are invited to apply for, for example, small business loans, or communities apply to participate in a program to provide water, schools or other social services) and administrative selection (the project implementing agency selects the individuals, communities or administrative areas who will participate). Hence there is very probably a selection bias as participants are likely to have special characteristics, often correlated with project success, which distinguish them from non-participants.

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

Where effect sizes are small, or there is a long time between treatment and measuring impact, then the problem of having to control for all confounding factors is greater, and so randomization again appears attractive. But the integrity of the design must be preserved, meaning that the control group must not receive any treatment.

The methodological advantage of RCTs is that they eliminate both project selection and sample selection bias—the two major threats to conclusion validity. This is a very powerful

advantage that cannot be matched by quasi-experimental designs. Some argue that regression discontinuity designs can also largely control for these biases, but this approach has been used infrequently in development evaluation, partly as the required longitudinal data sets are rarely available, but also because the enforcement of eligibility criteria (upon which the approach depends) is more likely to be lax in developing countries—an examination of the Grameen Bank microfinance program in Bangladesh found that, contrary to its stated criteria, many beneficiaries had more than one acre of land (Morduch, 1998). A study of a child growth monitoring program, also in Bangladesh, found that most community nutrition workers were unable to correctly interpret the growth charts used to select children for supplementary feeding (World Bank, 2005).

However, the advocacy for RCTs and strong evaluation designs, and the related, but separate, call for the creation of an independent evaluation agency, has stimulated a strong response from many critics, questioning the applicability, or the claimed advantages of these approaches. Most fundamentally, some critics argue that complex processes of social change cannot be assessed through quantitative outcome measures that ignore the setting within which the program is implemented, do not study how the implementation process affects outcomes, do not assess the qualitative dimensions of the program, and do not have the flexibility to identify and study changes that take place, both in project administration and in the project setting, during the life of the project. A powerful part of this critique, which has strongly influenced development agencies in the past, is the argument that intended outcomes such as 'empowerment' are immeasurable, certainly in terms of the money metric (see, for example, Kabeer, 1992). However, considerable progress has been made in the last decade in measurement of apparent elusive concepts such as social capital (e.g., Krishna, 2002) and, indeed, empowerment (Alkire, 2005, and Alsop et al., 2006). But even those who accept the validity of RCTs point out their limitations.

The first limitation is the narrow range of interventions to which RCTs can be applied. Such an approach is most readily applicable when the intervention is a discrete, homogenous intervention—precisely like a drug, and so the prevalence of the approach in the medical field, including in developing countries. But development projects are usually complex, heterogeneous and evolve during implementation. Examples where randomized approaches have been used are interventions such as conditional cash transfers (a grant to the household conditioned upon certain behavior, such as sending girls to school or ensuring children receive regular health check-ups), or very specific changes such as using flip charts in school or deworming school children. But for a large, multi-million dollar intervention, it is likely that only small sub-components of the project, if any, will be amenable to an experimental approach. For example, a single education project may support school rehabilitation, curriculum and textbook development, teacher training, strengthening local government management capacity, and technical support to the Education Management Information System—but at best selected sub-components of such a project will be amenable to a randomized evaluation. Development projects may also suffer from a small *n* problem, since the project may focus on working with one, or only a small number of agencies (such as supporting the creation of an anti-corruption commission or technical assistance to a government ministry), or support national policy reform.

A second limitation is that, even when an intervention appears to be homogenous, it can be difficult to ensure that this is in fact so across time and space. It is rarely possible to ensure a high degree of control over the conditions of subjects throughout the treatment period (which may last for a year or more in some

cases), or to ensure multiple repetitions of the treatment with different dosages on different subject groups. Often logistical problems make it difficult to ensure that school books, medicines or other inputs are delivered on-time and regularly re-supplied to all locations. This can be complicated by irregular electricity, computer networks or, in poorer countries, availability of fuel. In more than one project we have evaluated the project workers accompanying the evaluation team to the field were not recognized by the villagers, having not visited the village for up to two years, and in other cases the quality of the services may vary, when for example, health centers, schools or other facilities are understaffed, or not all staff receive the preparatory training or speak the local language. A major challenge for many evaluations is to be able to monitor or document what services have actually been received as project monitoring systems either do not provide this information (e.g. there is no record of how many patients receive both the full package of malaria treatment and are given the required orientation on the use of the kit) or the project administrators intentionally fail to report deficiencies (as when many teachers are absent during the planting or harvest seasons). Consequently, the evaluator often has only limited information on how uniformly treatments were actually administered.

Furthermore, when projects are administered in phases it will often be found that selection criteria will be modified as the characteristics of the target population are better understood or political pressures come into play to allow certain previously excluded groups to be considered (for example, a secondary school scholarship program may originally have been targeted for rural areas but children in urban areas may gradually be admitted). Alternatively, the package of services, or the way they are administered, may change.

Similarly, when the project is implemented in different locations, and particularly when these fall under different administrative jurisdictions, the selection procedures may vary—either because the nature of the data used to define the target population may differ or because of different political pressures. Or the project may use different implementing agents—sometimes government, sometimes non-governmental organizations (NGOs), often a number of them—in different parts of the country. So, for example, as mentioned in the Bangladesh example above, in a growth monitoring program the technical skills of community nutrition workers is crucial, but the ability of different NGOs to effectively train women for this role can vary enormously.

A further requirement is that the project environment remains constant for all subjects throughout the treatment period and that no external events interfere with the supposedly controlled implementation setting. It is often not possible to ensure this degree of control and external events such as the opening of a new factory, or the launch of a project by another agency might complement or interfere with outcomes of the project being studied. Or different project settings may be subjected to economic, political, demographic or other changes that may affect how the project is implemented in different locations.

In addition to difficulties in satisfying these requirements in most development contexts, there are a number of other problems affecting the utilization of RCTs. First, the RCT evaluation designs are by definition inflexible in that the same measurement instruments must be applied throughout the evaluation. This makes it difficult to capture or adapt to changing circumstances such as changes in participant selection procedures, how the project is administered or to contextual changes that might have a different effect on project and control groups. For example, urban renewal programs might affect the comparison groups, or the opening of new industries might affect the economic opportunities of the target population.

A further potential problem concerns "seepage" (or "contagion") when sectors of the control group become absorbed into the project population or when they have *de facto* access to project benefits. Examples of the latter include access to water supply by neighboring communities (which also reduces water availability for project households), non-eligible communities or individuals gain access to scholarship programs or other educational interventions, or information campaigns spread by word of mouth rather than project-supported media.

Another potential issue concerns the fidelity of the data collection procedures (how accurately they represent the situation being described. While this is an issue for all evaluation designs it may be particularly problematic for RCTs that rely on one or a limited number of quantitative indicators.

It is sometimes claimed that the implementation of an RCT design will be more expensive than a non-RCT. However, this will not always be the case as a randomized design will often significantly reduce the cost of sample frame construction as both project and control samples are generated from the same sampling frame.

Finally there may be practical difficulties to implementing a randomized design. The most common is that a random allocation is not acceptable to policy-makers, either because the program's scope is national, or where it is not, then administrative criteria are preferred or there is political interference in the allocation, which can also come into play once an intended experiment is underway. A project management office, which is typically a unit under a parent ministry, is not politically powerful, so higher-level political support is needed to experimental designs to enforce their implementation at local level.

A second practical problem is that evaluators are usually not involved in project design, and so potential opportunities to introduce strong evaluation designs are sometimes missed as the project design and the parameters of the evaluation have already been established before the evaluator is called in. Of course in some cases randomized designs are selected even without the involvement of the evaluator. One example is when demand for a particular service exceeds supply and a lottery is used to ensure that the system of selection is seen to be fair and to avoid political or other pressure during selection—but such cases are very rare in developing countries.

Third, there are sometimes ethical objections to the use of randomization as it is perceived that important benefits affecting health or even life-saving treatments are deliberately withheld from people who need them. Indeed, the term experiment has unfortunate connotations, as people may object to being 'experimented' on—randomized program designs for the Maori population in New Zealand were stopped for precisely this reason. The response to this, which may prove acceptable is that the control is not getting the treatment yet, rather than it is not getting it at all.

Finally, government policies or interventions of other agencies may differentially affect the project and control populations. In some cases government or other agencies may provides additional services to the target population to take advantage of the investments already being made by the project implementing agency. In other cases, these agencies may only focus on the non-beneficiary population to avoid duplication or to even out levels of benefits to different sectors of the low-income population.

The above considerations are just some of the constraints on using RCTs in developing countries. The fact that there have been so few such experiments is not simply because agencies have been unwilling or unable to implement such evaluations, but because it is often estimated that probably at most only about 5 percent of the total value of development finance is amenable to such an approach. Even

the most ardent proponents of RCTs in the development field admit that they "must necessarily remain a small fraction of all evaluations" (Duflo & Kremer, 2005, p. 205).

Despite these difficulties there are a number of situations in which randomized trials can be used. In principle the possibility of using an RCT can be considered whenever there is a clearly defined target population, subjects can be randomly allocated to treatment and control groups, a significant proportion of the population will not receive the treatment, the treatment is applied in a standardized and uniform way, and the project setting remains relatively stable throughout the period of the trials. Consequently, RCTs are likely to work better when the trial lasts for a relatively short period of time.

The following are some of the situations in which RCTs can be used. First, government may decide to use a lottery selection system to ensure equity and transparency in the selection procedure. The Bolivian Water Supply and Sanitation project was one example where a lottery was used because demand from villages interested in receiving these services exceeded the government's ability to provide the services within a given period (in this case programs were planned on an annual basis). The lottery was considered politically and ethically acceptable because this was a multi-year program so that villages not selected during the first year knew they had another chance to enter the lottery again the following year.

Second, there are a number of situations in which stakeholders are interested to test the effectiveness or cost-effectiveness of treatments in different combinations and settings and where randomization is possible and politically acceptable. These are often situations where replication of a program is being considered and where it is important to separate real project effects from other confounding factors. Examples where RCTs have been used include: comparing the effectiveness of deworming with other ways to increase school attendance or performance; the effectiveness of water supply and sanitation on community health and income and the comparison of the effectiveness of different delivery systems (community managed versus top-down); the impacts of vouchers for private schooling in Colombia; and the impact of changing interest rates on loan acceptance in South Africa.

Third, as Cook observes in his article, there are situations in which randomized trials are considered to be more credible than other options by some key stakeholders, so there may be political support for their use. Such an argument might apply to some countries in Latin America, where there is a strong tradition of impact evaluation (several countries require such studies to be carried out for all social programs), and randomized approaches have become widely known through their use for high-profile conditional cash transfer programs, notably PROGRESA in Mexico. But, outside of Latin America, such a systematic demand from policy makers is at present unlikely in developing countries.

Opportunities for Applying Strong Quasi-Experimental Designs in International Development

The problems of program heterogeneity and of a small *n* also limit the applicability of quasi-experimental impact evaluation designs. The problem of possibly immeasurable outcomes will hinder any quantitative approach. Nonetheless, although still only representing a minority of all project evaluations, the opportunities for applying strong quasi-experimental designs are much greater than for using RCTs. There are many situations where randomization was not possible, or was not used, but where a pretest-posttest control group design was used, which allows a double-difference based approach. As argued above, such approaches can be bias free unless there are unobserved determinants of program

participation which vary over time and which are correlated with the outcomes of interest. These designs can be used whenever survey data can be collected on both the project and a (reasonably representative) comparison group at the start and end (or at some point late in the project cycle) of the project. They are particularly strong when good secondary survey data are available so that propensity score matching or instrumental variables can strengthen the comparison between the two samples.

One of the main limitations on the use of strong quasi-experimental designs is that frequently the evaluation is not commissioned until late in the project cycle so that baseline data cannot be collected. This late start would of course also eliminate the possibility of using RCT. However, when the evaluation does begin at the start of the project, and if sufficient funds are available, it is often possible to use the pretest/posttest control group design. Examples include: evaluating housing and urban infrastructure projects targeted to clearly defined low income populations; conditional cash transfer programs; scholarship programs targeted for a particular section of the school population (for example female secondary students or all students from poor families when there is a clearly defined criterion of poverty); water supply and sanitation projects targeted for particular communities and road construction projects (although in these latter cases the definition of project and control groups is often more difficult, but not impossible to define).

Even if a formal baseline survey was not conducted, there may be other surveys which have been carried out in the project area which can serve this purpose. For example, in its evaluation of agricultural extension services in Kenya, the World Bank Independent Evaluation Group (IEG) commissioned a household survey of 285 households which had been covered by the Rural Household Budget Survey 15 years earlier, at the start of the National Extension Project (World Bank, 1999). In another example, for its study of support to basic education in Ghana, IEG surveyed 1,600 households and 705 schools in 85 enumeration areas which had been covered by a combined income and expenditure and education survey in the late 1980s (World Bank, 2003).

Despite the practical and political difficulties discussed above, strong evaluation designs have an important role to play in international development. Very few development programs have been subjected to rigorous impact evaluations, and the vast majority have been assessed without even a simple quasi-experimental design or any reference to a counterfactual. A large proportion of aid projects have not been subject to any impact evaluation, but even when there has been such a study it has not always employed a counterfactual. A review by the World Bank's evaluation department (OED, now called IEG) of its own impact evaluations found that of the 78 studies so classified only 21 had employed a counterfactual (Gupta Kapoor, 2002), though the situation has changed so that all new IEG impact studies contain counterfactual analysis. Consequently many of the claims that programs have been "effective" and have achieved their objectives (contributing to the elimination of poverty, increasing school enrolment and performance and so on) are often based on rather flimsy evidence. Many agencies define impact as simply comparing baseline measures with post project measures for the target population with no kind of comparison group and it is implicitly assumed that all of the changes can be attributed to the project intervention.

Recently there has been a renewed concern within the development community for greater aid accountability and many agencies have introduced results-based management, which it is claimed focuses on a better measure of results (outcomes and possibly impacts). However, in most cases the results-based management systems continue to rely on post project

comparisons with a baseline but with no comparison group, so it is not clear how much progress has been made.

Consequently, reasonably robust evaluation designs that include a logically defensible counterfactual, even if they do not satisfy the highest methodological standards, can provide a significantly better understanding of the extent to which development programs are achieving their objectives as well as helping understand the factors contributing to the level of impact and how it is distributed among different sectors of the target population.

As we will discuss in the next section, there are a wide range of impact evaluation design options that offer a useful level of understanding of potential impacts even when evaluations are conducted under budget and time constraints. Many of these techniques can also be used in the very common situation where the evaluation is not commissioned until late in the project cycle.

The main message is on the need to broaden the focus of the debate beyond searching for the relatively small number of project settings where rigorous designs can be used (although advantage should be taken of all opportunities to apply these designs), to focusing on proposing a wider range of impact evaluation methodologies than can provide operationally useful, and acceptably valid estimates, of project impact and that can be applied to a much wider range of development interventions. The challenge is to define a minimum set of methodological criteria for an evaluation design to be considered sufficiently rigorous to provide valid estimates of project impact.

Real-World Approaches to Impact Evaluation

Evaluations are commissioned either at the outset of a project (*ex ante*) or toward its end (*ex post*). The former case permits the use of stronger evaluation designs and much of the discussion on rigorous impact evaluations is limited to a discussion of this scenario. However, even when the evaluation is planned to commence with the start of the project there are a number of real-world constraints that limit the possibility of using strong evaluation designs (Bamberger, Rugh & Mabry, 2006). Budget constraints may exclude collection of baseline data for a comparison group and may even limit the kinds, and sample size, of baseline data which can be collected on the project population. In other cases time pressures may limit baseline data collection or make it impossible to conduct exploratory studies and pilot testing of data collection instruments required for a sound survey design. Time pressures may squeeze out the baseline altogether—once project implementation starts there is much to be done, and conducting a baseline survey for an impact study that will only produce results some years hence is far from a priority so it is eventually conducted, at best, a few years into the project (in a recent Indian irrigation project we studied, the 'baseline' had been conducted five and a half years into a seven year project). For this reason, it is preferable to conduct the baseline before the formal start of the project, which will usually require funds from a different source. It may also be difficult to identify populations not affected by the project from which a comparison group could be selected. Finally there may be political or administrative constraints such as the implementing agency's concern that interviewing families or communities not scheduled to receive project benefits will stir up political controversy or create pressures to expand the scope of the project beyond available plans or resources. In other cases the client is not convinced that it makes sense to "waste" money and time interviewing populations not involved in the project.

But it is more common that the evaluation is not commissioned until towards the end of the project, or even after it has finished—either

because funding and implementing agencies only become aware at this late stage of the need to collect systematic evidence on which to make decisions about continuation or replication of the project, or because the original project document required an impact evaluation but it was not considered a priority. Indeed, it is currently common practice amongst the evaluation departments of all major development agencies to not get involved in evaluation until the end of the project, although the project should also contain plans for a 'self-evaluation' implemented by the project staff. Only recently has the World Bank's evaluation department made an exception and allowed its staff to give *ex ante* advice on evaluation design for a health financing project in the Indian state of Karnataka. The French development agency has also recently begun an impact evaluation program with ex ante evaluation. But these examples remain exceptions.

Frequently, but not always, the belated interest in evaluation also means that there is an inadequate budget and often time pressures to deliver the evaluation report in time for the negotiations on the future of the project. Critics also claim that given that the evaluation is being commissioned to support the agency's claim that the project should continue to be funded, the evaluator is often given subtle, or not so subtle hints that while the evaluation must be "objective and impartial", it is hoped that the findings will be positive. However, our experience of working for a number of development agencies suggests this is not common practice, though the extent to which it happens varies and is more nuanced. In general evaluations undertaken for or by evaluation departments have a fair degree of independence. There are usually systems for review or response from the operational side of the agency, which can help correct errors of fact, though may sometimes also allow pressure to be brought to bear on content—though formal independence can limit these pressures. 'Self-evaluations' are more likely to be subject to these biases, but can still provide valuable lessons, though it may sometimes be necessary to read between the lines.

The range of possible quasi-experimental and non-experimental impact evaluation designs is summarized in Table 1. These approaches have been widely used in real-world contexts when experimental (randomized) designs have not been an option. The designs are ordered roughly from methodologically most to least robust. However, this is only a loose classification as theoretically sound designs can be considerably weakened if they are not properly implemented (which of course also applies to RCTs), while some of the theoretically weaker designs can be strengthened for example if used as part of a mixed-method, theory-based design or if additional observation points can be included.

It should also be emphasized that this categorization is made from a quantitative evaluation design perspective and many qualitative evaluation practitioners would take issue with the underlying premise of the superiority, or even the appropriateness of the quantitative methods on which the judgment is made.

Five of the designs (1, 2, 4, 5 and 7) can only be used when the evaluation begins at the start of the project. One design can be used when the evaluation begins when the project is already underway (design 3) and two are used for evaluations that start late in the project cycle (designs 6 and 8).

Strengthening the Evaluation Design

There are a number of cost-effective ways to strengthen impact evaluation designs when working under budget, time or data constraints. Indeed, these methods should be adopted for any impact evaluation. But they are particularly important when facing time or budget constraints as they help underpin the validity of the findings.

The first is to consider the feasibility of building the evaluation design on a program theory model. A theory-based approach involves mapping out the channels through which the inputs are expected to achieve the intended outcomes. When circumstances permit (see later in this paragraph), a program theory model helps explain the links in the causal chain enabling the evaluation to identify the key assumptions that must be tested. A program theory can also incorporate contextual analysis so as to identify local economic, political, institutional, environmental and socio-cultural factors that can help explain differences in the performance and outcomes of the same project when implemented in different locations. Theory-based approaches can also incorporate process analysis so as to monitor how the project is actually implemented, the quality of implementation and unplanned variations in the package of services actually received by different communities or beneficiaries. Under certain conditions the program theory can help distinguish between design failure and implementation failures to explain why intended outcomes were not achieved, and can help establish plausible association between inputs and outcomes—or the lack of such an association.

However, program theory models only work well under certain conditions. Theory models do not work well when there is no sound theory on which to build, or there is a lack of empirical evidence on, for example, expected effect sizes or the linkages between key variables. They are also difficult to use when there are several competing theories. However, Weiss (2000) argues that it is possible to develop and test several alternative theory models based on different theories. For example, Carvalho and White (2004) defined and tested two competing theories to explain the likely impacts of social investment funds on the level of local participation in the selection of community social infrastructure projects. One theory, advocated by the supporters of social funds, argued that inviting local communities to select among different social infrastructure projects would increase the level of community participation; while the other theory, espoused by some critics of the approach, argued that the decision-making process would be co-opted by local elites and would not increase local participation..

.A World Bank evaluation of agricultural extension services in Kenya found that extension workers spent far less time than planned in the field and visiting farmers, and that since the planned link from new research to extension advice did not operate, the extension workers were proposing to farmers that they adopt methods most had already adopted. Hence the result that there was no impact on yields is extremely plausible although the control was not a randomized one (World Bank, 1999).

A second method for strengthening evaluation design—for all evaluations not just weaker ones—is to adopt a good mixed-method design, combining quantitative and qualitative approaches in the formulation, implementation and analysis of the evaluation. This can be done in a number of ways. Qualitative data may be used for triangulation that is to provide additional evidence in support of the quantitative results. But the most important role for qualitative data is often to help frame the research. An evaluation design, and quantitative questionnaire, framed in ignorance of field conditions is very likely to overlook important aspects of how the project actually functions, which may well differ from what is described in the operational manual. Finally qualitative data can help interpret the quantitative results. A household survey conducted in Malawi and Zambia for a World Bank study of funds for community-identified and implemented projects (social funds) found that participation rates in the project selection decision meeting was very low, but participation rates in project implementation very high. Qualitative fieldwork showed that the decision on the choice of

Table 1
Eight Commonly Used Quasi-Experimental and Non-experimental Impact Evaluation Designs

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

Source: Adapted from Bamberger, Rugh, & Mabry (2006).

project was taken by a small group, usually the village headmen and the school head teacher, and then announced in the community meeting, with each household instructed to send a worker on a particular day (World Bank, 2002). Whilst this was not the community participation envisaged by the program's designers, it has proved an effective means of rapidly expanding social infrastructure in rural areas.

A third method is to make maximum use of available secondary data, including project monitoring data which are usually under-utilized in project evaluations. A fourth is to include, whenever time and budget permit, collection of data at additional points in the project cycle. In some cases this may be at some point during project implementation while in other cases this may involve data collection some time after the project has been completed so as to assess project sustainability.

Addressing Time and Budget Constraints

Addressing Budget Constraints

Five options can be considered (Bamberger, Rugh and Mabry. 2006. Chapter 3). First, considerable cost savings are often possible by eliminating one or more of the four data collection points (pretest/posttest project and control group). For example, design 5 eliminates baseline control group data and design 6 eliminates all baseline data. There is clearly a trade-off that must be assessed for this and the following options between cost savings and methodological rigor. Second, the data collection instruments can be simplified to reduce the amount of information to be collected. Often considerable amounts of unnecessary or low-priority information can be eliminated by judicious pruning. In other cases it may be possible to reduce the number of people from whom information is collected (for example, only interviewing the household head—although this can affect the quality of information on the opinions, behavior and economic activities of household members who are not interviewed). Third, the creative use of secondary data can often reduce data collection costs. Fourth, a judicious assessment of expected effect size and power analysis may sometimes make it possible to reduce sample size while still obtaining satisfactory estimates of project impact. Finally, there are often ways to reduce the costs of data collection. One possibility is to use less expensive interviewers such as medical students or student teachers rather than commercial interviewers. In some cases questionnaires could be self-administered (rather than hiring interviewers) and in other cases it may be possible to obtain information through direct observation rather than household surveys (for example, observing pedestrian and vehicular traffic patterns, or direct observation of time-use and sexual division of labor).

While it is often assumed that the evaluation will always require the collection of primary data, it is often possible to significantly reduce time and cost, as well as enhance quality by drawing on available secondary sources of data (White 2006). In addition to primary data collection in both project and control areas, it may be possible to obtain data from one of the following sources:

- *use of existing secondary data from already completed surveys* (demographic and health surveys, living standard measurement studies etc).*as a baseline for both project and control areas.*
- *use of secondary data, as discussed above, for control groups with the collection of primary data for project area.* This option is often used when the sample of project households in the secondary source is too small or where additional information, not included in the previous survey must be collected on the project population.

- *piggy-backing* in which an additional module can be added to an already planned survey, possibly over-sampling the project area to obtain the desired power
- *synchronized survey* in which a larger survey is used to select the control group (for example by propensity score matching) and a survey is carried out only amongst the project group.

The Inter-American Development Bank has been very successful in supporting low-cost impact evaluations while avoiding any new data collection. It has used proposals submitted to undertake studies to identify existing data sources, to which it can obtain access for local research teams who may not otherwise be able to obtain those data for analysis (and in consequence put in cheap bids to achieve this privilege).

Addressing Time Constraints

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

Addressing Data Constraints

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

First, as mentioned above, an existing survey may serve this purpose. Second, existing documentary data from within the organization or from other sources can be used, or key informants can also be asked to provide information on pre-project conditions. Finally, informants can be asked to recall their situation prior to the start of the project. Some evaluators question the validity of recall as it is particularly vulnerable to bias because of intentional distortion or lapses of memory. But all questionnaires are based on recall – so it is actually a question of degree rather than whether the approach should be used at all. Areas such as income and expenditure and fertility behavior, in which extensive research has been conducted on the reliability of recall, have shown that it is possible to identify the direction and magnitude of bias as well as identifying ways to reduce the bias. For example, between 1989 and 1998, the National Sample Survey in India experimented with different recall periods for measuring expenditure. It was found that when the 30-day recall period for food items was replaced with a 7-day period, the total estimated food expenditures increased by around 30%. When at the same time the 30-day recall period for infrequent expenditures was replaced with a one-year recall, the estimated total expenditure increased by about 17% (Deaton 2005). A number of studies have found there is a general

tendency to under-estimate small expenditures (truncation) and to over-estimate major expenditures (telescoping). Bamberger, Rugh and Mabry 2006 pp. 97-99 for a brief review of the recall literature.

Similar research in other areas (mainly by comparing information provided on current behavior or assets with recall of this same information at a future point in time) could greatly enhance the utility of recall. But for the time being it can be noted that major events and purchases (such as main assets like a vehicle or livestock) can be recalled with reasonable accuracy, especially if other methods are used to triangulate the information. Asset measures, combined with indicators of housing quality, are increasingly used as a proxy for the more difficult to measure outcome of household income. Krishna et al. (2005) use recall for an asset based approach to analyzing poverty trends in a number of Indian villages over a 25 year period.

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

Conclusions

We conclude that RCTs do indeed have a role to play in developing countries. Although, even under the most favorable circumstances RCTs will only make up a small percentage of impact evaluations, they are currently falling short of even that amount so there is scope for expansion, and given their limitations it is necessary to identify other means of undertaking impact studies. These other means must also address the time and budget constraints under which evaluators are frequently forced to operate. We have presented a range of designs with a range of costs and rigor. Where the most rigorous designs are not possible then a good theory-based approach will lend plausibility to the findings.

References

Alkire, S. (2002). *Valuing freedom: Sen's capability approach and poverty reduction*. Oxford: Oxford University Press.

Alsop, R., Bertelsen, M., & Holland, J. (2006). *Empowerment in practice: from analysis to implementation*. Washington, DC: World Bank.

Bamberger, M., Rugh, J., & Mabry, L. (2006). *Realworld evaluation: Working under budget, time, data and political constraints*. Thousand Oaks, CA: Sage.

Bamberger, M. (2006). *Conducting quality impact evaluations under budget, time and data constraints*. Washington DC: World Bank.

Carvalho, S & White, H (2004) "Theory-based Evaluation: the Case of Social Funds" American Journal of Evaluation 25(2):141-60.

Chambers, R. (1994a) "The origins and practice of participatory rural appraisal" *World Development* **22**(7): 953-969.

Chambers, R. (1994b) "Participatory rural appraisal: analysis of experience" *World Development* **22**(7): 1253-1268.

Chambers, R. (1994c) "Participatory rural appraisal: challenges, potentials and paradigm" *World Development* **22**(7): 1437-145.

CGD. (2006). *When will we ever learn? Improving lives through impact evaluation*. Washington, DC: Center for Global Development.

Cook, T. D. (2006) Describing what is special about the role of experiments in contemporary educational research: Putting the "gold standard" rhetoric into perspective. *Journal of MultiDisciplinary Evaluation*, *6*, 1-7.

Davidson, E. J. (2006). The RCTs-only doctrine: Brakes on the acquisition of knowledge? *Journal of MultiDisciplinary Evaluation*, *6*, ii-v.

Deaton, A. (2005) "Measuring Poverty in a Growing World (or Measuring Growth in a Poor World)" Review of Economics and Statistics 87(1):1-19

Duflo, E., & Kremer, M. (2005). Use of randomization in the evaluation of development effectiveness. In G. K. Pitman, O. Feinstein, & G. Ingram (Eds.), *Evaluating development effectiveness*. Washington, DC: World Bank.

Gupta Kapoor, A. (2002). *Review of impact evaluation methodologies used by the Operations Evaluation Department over past 25 years* [OED Working Paper]. Washington, DC: IEG, World Bank.

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

Krishna, A. (2002). *Active social capital: Tracing the roots of development and democracy*. University Presses of California, Columbia and Princeton

Krishna, A., Kapila, M., Porwal, M., & Singh, V. (2005). Why growth is not enough: Household poverty dynamics in Northeast Gujarat, India. *Journal of Development Studies*, *41*(7), 1163-1192.

Kumar, S. (2002). *Methods for community participation: A complete guide for practitioners*. Rugby, England: ITDG Publishing.

Morduch, J. (1998). *Does microfinance really help the poor? New evidence from flagship program in Bangladesh*. Princeton, NJ: Princeton University.

Scriven, M. (2006). Converting perspective to practice. *Journal of MultiDisciplinary Evaluation*, *6*, 8-9.

Weiss, C (2000). "Which Links in which Theories Shall We Evaluate?" Pp 35-45 in Program Theory in Evaluation: Challenges and Opportunities edited by P.J Rogers et al. New Directions for Evaluation. No. 87 San Francisco. Jossey-Bass.

White, H. (2004). Using development goals and targets for donor agency performance measurement. In R. Black & H. White (Eds.), *Targeting development: Critical perspectives*

on the Millennium Development Goals. London: Routledge.
White, H. (2006). *Impact evaluation experience of the Independent Evaluation Group of the World Bank*. Washington, DC: World Bank.
World Bank. (1999). *Agricultural extension: The Kenya experience*. Washington, DC: World Bank.
World Bank. (2002). *Social funds: Assessing the effectiveness*. Washington, DC: World Bank.
World Bank. (2004). *Books, buildings and learning outcomes: An impact evaluation of World Bank assistance to basic education in Ghana*. Washington, DC: World Bank.
World Bank. (2005). *Maintaining momentum to 2015? An impact evaluation of interventions to improve maternal and child health and nutrition outcomes in Bangladesh*. Washington, DC: World Bank.

Quasi-experimental design involves developing a comparison groups by using matching and reflexive techniques. Under such design, intervention sites are compared with non-intervention sites by using statistical methods to account for differences between the two groups. Under non-experimental design, intervention sites are compared with non-intervention sites by using statistical methods to account for differences between the two groups.Â 16 Bamberger, M. and H. White. 2007. Using Strong Evaluation Designs in Developing Countries: Experience and Challenges. Journal of MultiDisciplinary Evaluation. Vol 4(8):58â€“73. 17 Vaessen, J. and D. Todd. 2007. Methodological challenges in impact evaluation: The Case of the Global Environment Facility. Using strong evaluation designs in developing countries: Experience and challenges. Journal of Multidisciplinary Evaluation 4(8), 58â€“73. Berger, P., & Luckmann, T. (1966).Â Teachers views of school psychologists in different countries. International School Psychology Association â€“ World Go Round 27(5):8â€“9. Fixsen, D., Blase, K., Naoom, S., & Wallace, F. (2009).