WPS5245

Policy Research Working Paper 5245

Using Mixed Methods in Monitoring and Evaluation

Experiences from International Development

Michael Bamberger
Vijayendra Rao
Michael Woolcock

The World Bank
Development Research Group
Poverty and Inequality Team
March 2010

Policy Research Working Paper 5245

Abstract

This paper provides an overview of the various ways in which mixing qualitative and quantitative methods could add value to monitoring and evaluating development projects. In particular it examines how qualitative methods could address some of the limitations of randomized trials and other quantitative impact evaluation methods; it also explores the importance of examining "process" in addition to "impact", distinguishing design from implementation failures, and the value of mixed methods in the real-time monitoring of projects. It concludes by suggesting topics for future research—including the use of mixed methods in constructing counterfactuals, and in conducting reasonable evaluations within severe time and budget constraints.

This paper—a product of the Poverty and Inequality Team, Development Research Group—is part of a larger effort in the department to integrate qualitative and quantitative methods for monitoring and evaluation. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at vrao@worldbank.org.

Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development[1]

Michael Bamberger, Independent Consultant
Vijayendra Rao, World Bank
Michael Woolcock, World Bank and University of Manchester

I. Introduction

In recent years there have been increasing demands to measure the effectiveness of international development projects. This demand has emerged in response to two concerns: (1) Heightened criticism that most development agencies report only their outputs (e.g., number of teachers trained, kilometers of roads built) rather than outcomes; and (2) Concerns that, despite assurances that development resources have contributed to the reduction of illiteracy and poverty, little reliable information has actually been presented to show that this is in fact the case.[2] This has led to a demand for more effective and innovative ways to monitor and evaluate the nature and extent of impacts stemming from development initiatives (Rao and Woolcock 2003).

Quantitatively oriented development researchers, if not the development community as a whole, have responded enthusiastically to the evaluation challenge. In the last decade there has been an explosion of quantitative impact evaluations of program interventions in international development. This has been driven both by trends within academia and pressure from international organizations like the World Bank, and has culminated in efforts to adopt the standards and methods of bio-medical clinical trials in making knowledge claims about the effectiveness of particular interventions. Some development economists (e.g., Banerjee 2007, Duflo and Kremer 2005) have gone as far as to argue that randomized trials should be central to development practice, and that knowledge claims based on alternative approaches are not merely inferior, but inherently suspect. Others (Deaton 2009, Ravallion 2008), while accepting the central importance of quantitative evaluations, question the exclusive reliance placed by the "randomistas" on randomization and make the case that careful theorizing and tests of hypothesis that derive from theory, along with other types of quantitative methodologies – propensity score matching, careful structural modeling, and instrumental variables – should be not be so easily dismissed. The central concern of all these quantitative techniques is with obtaining statistically rigorous counterfactuals.

---

[1] The views expressed in this paper are those of the authors alone, and should be attributed to the respective organizations with which they are affiliated. Email addresses for correspondence: jmichaelbamberger@gmail.com, vrao@worldbank.org, and mwoolcock@worldbank.org. We thank the Handbook's editors for helpful comments, and our respective collaborators for all that they have taught us on these complex issues over many years.

[2] The absence of evidence cuts both ways of course: critics also have little empirical basis on which to claim that development projects unambiguously *haven't* worked.

The strong claims by quantitative development researchers to have the best (the "gold standard"), or even the only acceptable way to evaluate the effectiveness of development assistance has created a strong reaction from other sectors of the development community.[3] While some argue that randomization is rarely possible in developing countries, that the rigorous designs introduce an inflexibility that makes it difficult to understand the complex environment in which development projects operate, or that there are fundamental methodological issues with the approaches, many take the view that randomized and strong statistical designs are only one of a number of approaches to development evaluation. For example, NONIE (the Network of Networks on Impact Evaluation[4]) recently published "NONIE Guidance on Impact Evaluation" (April 2009) which recognized the importance of rigorous quantitative methods for addressing issues of causal attribution but recommended that evaluators should use a mixed method approach that combines the strengths of a range of quantitative and qualitative methods. In the latest edition of his *Utilization Focused Evaluation*, Patton (2008) lays out an extensive menu of different evaluation designs that can be used to address different kinds of evaluation questions.

None of the quantitatively oriented development participants in this debate, however, question the singularity of the econometric analysis of survey-based data as the core method of relevance for impact evaluations. Outside the context of impact evaluations, most work by economists on development questions also remains entirely econometric, though there is an increasing (and welcome) trend towards direct engagement in fieldwork and having survey data be analyzed by the person(s) who collected it, which in turn has led to a deeper and richer understanding of development problems. However, given that a central challenge in international development is that the decision makers (development economists included) are in the business of studying people separated from themselves by vast distances – social, economic, political and geographic – there is a strong case for using mixed methods to both help close this distance and to more accurately discern *how* outcomes (positive, negative, or indifferent) are obtained, and how any such outcomes vary over time and space (context).

By restricting themselves to the econometric analysis of survey data, development economists are boxed into a Cartesian trap: the questions they ask are constrained by the limitations inherent in the process by which quantitative data from closed-ended questions in surveys are collected (Rao and Woolcock 2003). As such, they are limited in their ability to ask important questions about the social, cultural and political context within which development problems are embedded. They even miss important aspects of some critical economic issues such as, for instance, the heterogeneity that underlies what are known as "informal" economies (i.e., labor markets that function outside formal salary and wage structures) and tend to overlook marginal markets that are centrally important for policy – such as the market for drugs, political favors, and sex – all of which require a strong degree of rapport with respondents that a short visit to field a questionnaire will not provide. A related criticism (to which we return later) is that many kinds of econometric analysis fail to examine what actually happens during the process of project implementation (the "black box") and consequently are unable to determine the extent to which failure to achieve intended impacts is due to "design failure" or to "implementation failure". In

[3] See also the critique of philosophers, such as Cartwright (2007).
[4] NONIE is supported by and brings together the DAC Evaluation Network, the Evaluation Cooperation Group of the multilateral finance institutions, the International Organization for Cooperation in Evaluation, and the UN Evaluation Group

other words, their research questions are being shaped by their data instead of their data by the questions. A strong case can be made that such questions require a more eclectic approach to data, one that mixes participation, observation, the analysis of text-based information (e.g., from tape recordings of village meetings or newspaper articles), free-ranging open-ended interviews with key informants and focus groups, and other such types of information that can be loosely called "qualitative data". The name is a bit of a misnomer, however, since a lot of qualitative data can be coded, quantified and econometrically analyzed. This distinction should therefore be more appropriately between data collected from structured, closed-ended questions and non-structured, open-ended, modes of enquiry.

Another argument for using mixed methods is that most program assessments have focused on assessing the tangible changes that can be measured over the 3-5 year life of most projects and programs funded by international (and national) development agencies. While there are a few large-scale and high profile impact evaluations that are cited in the literature, most attempts to assess impacts are based on evaluations that are not commissioned until late in the project and that have to be conducted over a short period of time and on a relatively modest budget. This is far from ideal, and concerted efforts need to be made to build impact evaluations into projects from their inception, but it is the reality that evaluators must confront and to which they should be expected to be able to make a constructive and valid contribution; mixed methods can be a useful approach to making such a contribution. Finally many evaluations, particularly those that are non-quantitative, have a systematic positive bias (Bamberger 2009b) because the short period that consultants are in the field frequently means that only project beneficiaries and agencies directly involved in project implementation are interviewed and most of these tend to have a favorable impression of the project (as they are the people who have benefited directly). As such, many project evaluations do not interview anyone who is not a beneficiary or at least involved in the project. Employing a mixed methods design can help to more accurately identify comparable 'non-participant' locations and individuals, and specify what a plausible 'counterfactual' – that is, what might have happened to participants were it not for the project – might look like in a given case (Morgan and Winship 2007).

Finally, there is increasing acceptance of the idea that monitoring is as central to good development practice as evaluation. Evaluations tell us how well a project worked; monitoring provides real time feedback, allowing the project to learn by doing and to adjust design to ground-level realities. A well designed monitoring system can often strengthen the baseline data required for pretest-posttest evaluation design, in some cases complementing a baseline survey by adding more project-related data and in others providing baseline data when the late commissioning of the evaluation did not permit the use of a baseline survey. While evaluations are now broadly perceived as central to development practice and have received a lot of intellectual attention, scholars and practitioners have paid less attention to monitoring methods. Here again mixed-methods can provide the flexibility to integrate good monitoring practices within evaluation designs.

The response to these contending pressures has been a very lively but as yet unresolved debate on the optimal way to monitor and evaluate the impacts of development assistance. Many aspects of these debates are philosophical and are couched in terms that are inherently contested (and thus will never have a neat resolution), but we will argue in this chapter that mixed method

approaches can nonetheless contribute to these debates in substantively important ways. The chapter proceeds as follows. Section II explores five key issues at the center of debates pertaining to project evaluation in developing countries[5]. Section III moves from 'first principles' to pragmatism, stressing how evaluators can use mixed methods in the less-than-ideal circumstances they are likely to encounter in the field, and especially in developing countries. Section IV focuses on the importance of understanding project processes and contexts, and building effective monitoring systems. Section V considers some specific way in which mixed methods can improve qualitative evaluations. Specific case studies of mixed methods evaluation in international development programs are provided in Section VI. Section VII concludes by considering the ongoing challenges and opportunities for using mixed methods in project evaluation in developing countries.

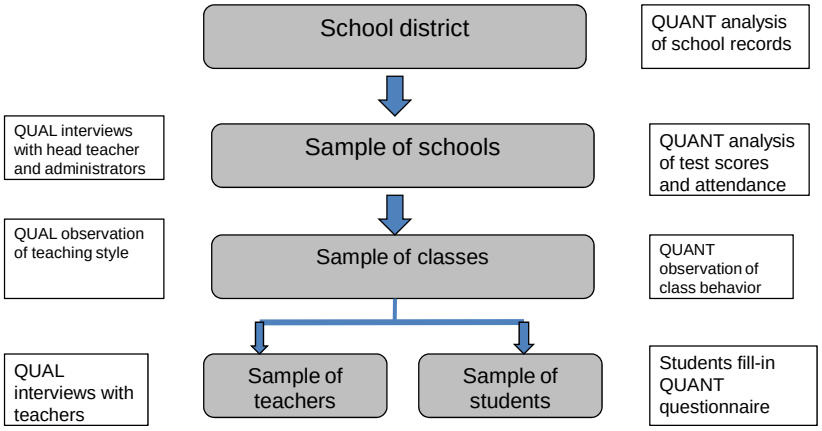
II. Using Mixed Methods in Project Evaluation and Monitoring: Five Key Issues

Among the many issues that confront program evaluators, at least five speak to the importance of using mixed methods. The first is the principle that the questions being posed should determine the research methods being used, not the other way around. As noted above, while most economists argue that randomized control trials (with strong quasi-experimental designs as a second best)[6] should be considered the "gold standard" for impact evaluation (see next point), two counter-arguments can be considered. One is that there are many different kinds of development assistance ranging in complexity and scope from the kinds of clearly defined projects for which conventional impact evaluation models were designed, to complex, multi-component national level programs that involve many donors, numerous national agencies and line ministries, and which often have no clear definition of their scope or intended outcomes and impacts[7]. We argue that the great diversity of projects requires the use of a range of correspondingly different evaluation methodologies, and specifically that strong statistical project level designs are often inappropriate for the more complex, multi-component programs. We recognize, however, that certain *aspects* of these complex programs can usefully be subjected to randomized designs (see, for example, Olken 2007). The second counter-argument is that evaluations are conducted to address a wide range of operational and policy questions – i.e., not just average treatment effects – and that, as such, different methodologies are required depending on the specific information needs of particular clients[8].

[5] All of these issues are also relevant to the application of mixed methods evaluations in industrialized countries.
[6] We will refer to RCT and strong quasi-experimental designs as *strong statistical designs*.
[7] Some of the dimensions that can be used for classifying evaluation scenarios include: (1) Evaluation purpose, (2) Level at which the evaluation is conducted, (3) Program size or scale, (4) Complexity of the evaluand, (5) size of the evaluation budget, (6) Stage of the project at which the evaluation is commissioned, (7) Duration of the evaluation, (8) Who is the client, (9) Who will conduct the evaluation, (10) level of required statistical rigor, (11) Location of the evaluation design on the QUANT/QUAL continuum, and (12) Source of data (primary, secondary, both).
[8] Examples of evaluation purpose include: (1) Generalizability (assessing the feasibility of replicating or scaling up an intervention, (2) Developmental (tracking emergent, complex interventions), (3) Accountability (Ensuring accountability for results), (4) Contribution/substitution (assessing the contribution of different donors to comprehensive, collaborative interventions, or assessing with donor funding produces a net increase in funding for a program or whether this replaces previously allocated government funding). Source: Adapted from NONIE 2008.

Figure 1 illustrates how a mixed method approach can be used for multiple-level analysis of a complex education program that must be assessed simultaneously at the level of the school district, the school, the classroom and the individual student or teacher. Similar multi-level analysis could be used for the analysis of multi-component national level programs.

Figure 1 Multi-level nested mixed methods design:
Evaluating effects of school reforms on student attendance and performance



One of the emerging questions is whether there are alternatives to the statistical counterfactual that can be used in attribution analysis when it is not possible to use a quasi-experimental design with a well-matched comparison group. Most of the alternatives to the strong statistical designs rely on mixed method approaches that often combine case studies of households, communities or organizations with key informant interviews and/or focus groups, with the synthesis of available quantitative data sources being used to ensure that the qualitative data is adequately representative and to permit the extrapolation of the findings to other populations. The development of valid alternatives to the statistical counterfactual is still a new and emerging field, one that usually relies on a combination of methods to "reconstruct" a hypothetical scenario absent the project intervention. Some of the techniques that are used include: using two or more program theory models to compare expected outcomes under the intended project design and alternative hypotheses; participatory techniques such as PRA to obtain beneficiary perceptions on causality; concept mapping (Kane and Trochim 2007); and techniques such as Theory of Change (Morra and Rist 2009: 150-169). All of these approaches rely on combining different types of data collection and analysis. This field is still in its infancy, but to the extent that strong statistical designs can only be used in a small minority of evaluations, is a potentially very important area for which mixed method approaches are well suited. Assessing how well these methods work in comparison to randomized impact evaluations is an important area of future research.

The second issue concerns the application of strong statistical designs in development contexts. We will not revisit the ongoing debate about the technical, ethical, political and practical merits of RCTs and other statistical designs, but will only consider contexts in which it has already been agreed to use one or other of these designs. While these designs have significant statistical advantages in terms of the elimination or reduction of project selection and sample selection bias, when used on their own these designs have a number of fundamental weaknesses. We argue that mixed methods can significantly strengthen the validity and operational utility of these designs, as well as the design and implementation of more traditional quantitative evaluation strategies using double difference and matching strategies. Indeed, we argue that "rigor" is not determined solely by the use of a particular method as such, but rather the appropriateness of the 'fit' between the nature of the problem being assessed and the particular methods (singular or in combination) deployed in response to it, given prevailing time, political, financial, ethical and logistical constraints.

The following are some important ways in which mixed methods can contribute to further strengthening these designs:

(a) Most strong quantitative impact evaluation designs [strong designs for short] use a pre-post test comparison design. Information is usually not collected on the process of project implementation. Understanding "process" is the central concern of monitoring and, we would argue that effective monitoring is necessary for effective evaluation. For instance, if the analysis does not find any statistically significant differences between the project and comparison groups, it is not possible to determine whether this is due to *design failure* (the proposed design is not appropriate to achieve the intended objectives in the particular context) or to *implementation failure* (the project was not implemented as planned so it is not possible to assess the validity of the project design or to recommend whether the project should be replicated). A mixed methods approach could incorporate process analysis through the use of qualitative techniques such as participant observation, key informant interviews and focus groups to assess the process of project implementation and how this affected program outcomes and impacts. Similarly, the absence of process data precludes analysis of the nature and extent of impact trajectories—that is, whether and how a project is performing not just with respect to a counterfactual, but with respect to what theory, evidence or experience would suggest that particular *type* of project should be achieving, given how long it has been implemented (Woolcock 2009). Most evaluations, for example, assume that project impact is monotonic and linear, whereas experience alone would suggest that, for even the most carefully designed and faithfully implemented project, this assumption is not often valid.

(b) Most traditional evaluation designs rely on a limited number of uni-dimensional quantitative indicators to measure project impacts and outcomes. However, many constructs (such as poverty, empowerment, community organization and leadership) are complex and multidimensional, rendering conventional quantitative indicators vulnerable to construct validity issues. Mixed methods can contribute a range of qualitative indicators as well as generating case studies and in-depth interviews to help understand the meaning of the statistical indicators. These indicators can sometimes be aggregated into quantitative measures and thereby incorporated into the statistical analysis.

(c) Most quantitative data collection methods are not appropriate for collecting information on sensitive topics (such as domestic violence, operation of community and other kinds of organizations, social and cultural factors limiting access to services), or for locating and interviewing difficult-to-reach groups (e.g., crime bosses, sex workers, members of marginalized social groups). There are a range of mixed methods techniques that can help address these issues.

(d) Strong designs are usually inflexible in that the same data collection instrument, measuring the same indicators must be applied to the same (or an equivalent) sample before and after the project has been implemented. This makes them much less effective for real-time learning by doing, and for monitoring. Projects are almost never implemented exactly as planned (Mosse 2005) and consequently the evaluation design must have the flexibility to adjust to changes in, for example: project treatments and how they are implemented; definition of the target population; and changes in the composition of the control group. Mixed methods can provide a number of rapid feedback techniques to provide this flexibility to adapt to changing circumstances.

(e) Strong designs are also criticized for ignoring (or being unable to incorporate the range of) the local contexts in which each project is implemented, which in turn can produce significant differences in the outcomes of projects in different locations. Mixed methods can help provide detailed contextual analysis. A well-designed monitoring system can document differences in the quality or speed of implementation in different project locations that might suggest the operation of local contextual factors. However, the analysis of these factors will normally require broadening the scope of conventional monitoring systems by the incorporation of some of the mixed method data collection and analysis techniques discussed in Section IV.

(f) Similarly, RTCs and other quantitative evaluation methods designed to estimate the average treatment effects do not capture heterogeneity in the treatment effect (Deaton 2009). While there are useful quantitative techniques designed to deal with treatment heterogeneity, qualitative methods can also be a strong aid to understanding how the treatment may have varied across the target population.

(g) Sample selection for strong designs is often based on the use of existing sample frames that were developed for administrative purposes (such as determining eligibility for targeted government programs). In many cases these sampling frames exclude significant sectors of the population of interest, usually without this being recognized. This is particularly true of regression discontinuity designs that necessarily exclude target populations outside the range of the discontinuity within which the treatment effect is identified. Mixed methods can strengthen sample coverage through a number of techniques such as on-the-ground surveys in selected small areas to help identify people or units that have been excluded.

The third set of issues concern adapting impact evaluation to the real-world contexts and constraints under which most evaluations are conducted in developing countries. While evaluation textbooks provide extensive treatment of rigorous impact evaluation designs – that is, designs that can be conducted in situations where the evaluator has an adequate budget, a reasonable amount of time to design, conduct and analyze the evaluation findings, access to most of the essential data, and a “reasonable” degree of political and organization support – most

textbooks offer very little advice on how to conduct a methodologically robust impact evaluation when one or more of these conditions do not obtain. Most of the recent debates have focused on advocating or criticizing the use of RCTs and strong quasi-experimental designs, and there has been almost no discussion in the literature on how mixed methods can help improve the conduct of impact evaluations in the vast majority of cases where rigorous statistical designs cannot be employed[9]. For many evaluation professionals, particularly those working in developing countries, the debates on the merits and limitations of statistically strong impact evaluation designs are of no more than academic interest as many may never (and are highly unlikely to) have an opportunity to apply any of these designs during their whole professional career[10]. (In the following section we will discuss how mixed methods evaluation can help adapt evaluation theory to the real-world time, budget, data and political constraints under which most evaluations are conducted in developing countries.)

A fourth set of issues reflect the widespread concern about the low rate of evaluation utilization (Patton 2008). Many methodologically sound evaluations are not used or do not contribute to the kinds of changes for which they were commissioned. There are many reasons for this (Bamberger, Mackay and Ooi 2004, 2005; Pritchett 2002), but high among them is the fact that evaluation designs often do not have the flexibility to respond to the priority questions of concern to stakeholders, or that the findings were not available when required or did not use a communication style with which clients were comfortable. Often the evaluation paradigm (whether quantitative or qualitative) was not accepted by some of the stakeholders. While mixed method approaches do not offer a panacea for all of the factors affecting the utilization of evaluation, they can mitigate some of the common causes of low uptake. For example, purely quantitative designs – with their requirement for standardized indicators, samples and data collection instruments – usually do not have the flexibility to respond to the wide range of questions from different stakeholders, whereas qualitative techniques usually do have this flexibility. Mixed methods can also incorporate process analysis into a conventional pre-test/post-test design, thus making it possible to provide regular feedback on issues arising during implementation and to provide initial indications of potential outcomes throughout the life of the project, hence increasing the ability to provide timely response to stakeholder information needs. The collection of both quantitative and qualitative information also makes it possible to use different communication styles for presenting the findings to different audiences. Reports on case studies, perhaps complemented by videos or photographs, can respond to the needs of audiences preferring a human interest focus, while the statistical data can be presented in tables and charts to other more quantitatively oriented stakeholders.

III. Using Mixed Methods to Conduct Evaluations under Real-World Constraints

When actually conducting evaluations of development projects, one frequently faces one or more of the following political and institutional constraints, all of which affect the ability to design and

[9] Although no formal statistics are available, the present authors would estimate that pretest/posttest comparison group designs are used in less than 25% of impact evaluations, and quite probably in less than 10% of cases.

[10] One of the authors organizes an annual workshop (International Program for Development Evaluation Training) for 30-40 experienced evaluation professionals working in developing countries. Every year he polls them on how many have been involved in a strong statistical evaluation design, and often not a single participant has ever had the opportunity to use one of these design.

implement “rigorous” impact evaluations. First, many evaluations are conducted on a tight budget which limits the ability to carefully develop and test data collection instruments, and which often makes it impossible to use the sample sizes that would be required to detect statistically significant impacts, particularly for the many situations in which even well-designed projects can only be expected to produce a small change.

Second, many evaluations are conducted with a very tight deadline which limits the time available for data collection and often the amount of time that consultants (many of whom are expensive foreigners) can spend in the field. Another dimension of the time constraint is that evaluators are often asked to assess outcomes or impacts when it is too early in the project cycle to be able to obtain such estimates. A less discussed but equally salient constraint is the limited time that clients and other stakeholders are able, or willing, to discuss the evaluation design or the preliminary findings. Consequently, if program theory models are developed there will often be very little input from clients, thereby defeating the intention of ensuring that stakeholders are fully engaged in the definition of the program model that is being tested and the output and impact indicators that are being measured

Third, many evaluations have very limited access to the kinds of data required for constructing baselines or comparison groups. This is particularly problematic in those cases (probably the majority) where the evaluation is not commissioned until late in the project cycle. Frequently no baseline data has been collected on the project group and no attempt was made to identify a plausible comparison group. Sometimes project administrative records or surveys designed for other purposes may provide some data but often it does not cover the right population, contain the right questions, collect information from the right people (only the “household head” but not the spouse or other important household members), or refer to the right time period. In other cases the information is incomplete or of a questionable quality. Another dimension of the data constraint is when the data collection instruments are not well suited for obtaining information on sensitive topics (sexual practices, illegal/illicit sources of income, corruption or domestic violence), or for identifying and interviewing difficult-to-reach groups (e.g., those who are HIV positive, drug users, illegal immigrants).

The fourth, and final, set of constraints relate to political and organizational pressures that affect how the evaluation is formulated, designed, implemented, analyzed and disseminated. These are all influenced, for example, by what issues are studied (and not studied), which groups are interviewed (and not interviewed), what kinds of information are made available (and withheld), who sees the draft report and is asked to comment (and who does not), and how the findings are disseminated. The pressures can range from a subtle hint that “this is not the time to rock the boat”, to a categorical instruction from public authorities that it will not be possible to interview families or communities that have not benefited from the project.

The combined influence of these factors can have a major influence on the type of design that can be used, the sample sizes, the use of statistical significance tests and the levels of methodological rigor (validity) that can be achieved. A frequent problem is that the tight budget and the short timeframe that consultants can spend in the field is often considered as a justification for only visiting project sites and only interviewing beneficiaries and stakeholders directly involved in (and usually benefiting from) the project. As a result, many evaluations

have a systematic positive bias because they only obtain information from beneficiaries. As these rapid (but flawed) evaluation approaches show funding and implementing agencies in a good light, there has been much less concern about the widespread use of these fundamentally flawed methodologies than might have been expected.

Other common problems include: the sample sizes being too small to be able to detect impacts even when they do exist; techniques such as focus groups are often seen as a "quick and dirty" way to capture community or group opinions when there is neither time nor money to conduct sample surveys and accepted practices for selection of subjects or avoiding bias in data collection and analysis are ignored; attribution is based on comparisons with poorly selected and not very representative control groups; data is only collected from one household member (often whoever is available) even when the study requires that information and opinions are collected from several household members.

Many quantitative evaluation designs are commissioned late in the project cycle and rely on secondary data from previous surveys for the baseline comparison. However, as the secondary surveys were normally commissioned for a different purpose than the project evaluation (national income and expenditure panel survey, national health or nutrition survey) it will often be found that the surveys do not contain all of the required information (particularly concerning specific information on access to the project), do not collect information from the requisite people, were not conducted at the right time or do not completely cover the target population. While statistical techniques such as propensity score matching and instrumental variables can help strengthen the match of the project and comparison groups and eliminate some kinds of selection bias, there are a number of challenges that normally cannot be resolved. In addition to the above problems of sample coverage, who was interviewed etc, a major challenge frequently concerns missing information on differences between the project and comparison populations that might explain some of the post-project differences (e.g., in income, school test performance, infant mortality) that are assumed to have been produced (at least in part) by the project. For example, the women who apply for small business loans may mainly come from the small group of women who have the self-confidence and experience to start a business, or who have an unusually high degree of control over household decision-making or who have husbands supportive of their economic independence. Most applied econometric analysis tries to downplay the importance of these "unobservables" by assuming they are "time invariant" and consequently can be differenced out by using panel data. However, this assumption is often highly questionable, particularly when the researcher may have no idea what these factors might be.

A final challenge, one that will be discussed in more detail in the following section, concerns the lack of information on what happened during the project implementation process. Projects are rarely implemented according to plan, and often access to services and the quality of the services will be significantly less than planned. Consequently, in explaining why intended outcomes and impacts were not achieved, it is essential to know whether lack of results was due to design failure (the basic design did not work in this context) or whether problems were due to implementation failure. In real-world evaluation contexts the information is often not available to make this judgment.

How can mixed methods address these real-world challenges?

Mixed method approaches combine quantitative approaches that permit estimates of magnitude and distribution of effects, generalization and tests of statistical differences with qualitative approaches that permit in-depth description, analysis of processes and patterns of social interaction. These integrated approaches provide the flexibility to fill in gaps in the available information, to use triangulation to strengthen the validity of estimates, and to provide different perspectives on complex, multi-dimensional phenomena. When working under real-world constraints, a well designed mixed-methods approach can use the available time and resources to maximize the range and validity of information. The following are some of the specific ways that mixed methods can strengthen the impact evaluation design when working with real-world constraints.

Reconstructing baseline data. There are a number of qualitative techniques that can be used to reconstruct baseline data. These include: recall, key informants, participatory group interview techniques (such as PRA), and the analysis of administrative records (such as school attendance, patient records from health clinics, sales records from agricultural markets) (see Bamberger 2009a). Ideally a triangulation strategy should be used to increase validity by comparing estimates from different sources. Most of the above sources use purposive, often opportunistic, sampling methods relying on information that is easily accessible, and consequently there are questions of bias or non-representativity. A mixed method strategy can include statistical methods either to use random sampling techniques (for example to select the schools or clinics) or to assess the potential direction and magnitude of bias.

A well-designed monitoring system can often provide at least some of the baseline data required for impact evaluation (see following section). However, experience has shown that the generation of data in the format required for impact analysis will usually require the active participation of the evaluator in the design of the monitoring system, both to ensure that the required information is collected and also to ensure that it is organized and filed in a way that will be suitable for the evaluation. Often potentially valuable monitoring data has proved impossible to use for evaluation because certain critical classification data (such as code numbers identifying beneficiaries, types of services received, reasons for dropping out of the project etc.) have not been included or recorded systematically. In other cases the information is incomplete or some has been mislaid because staff did not realize that this information was intended for use in the evaluation. Sometimes the quality of the information could have been greatly enhanced if the evaluation budget could have funded additional administrative staff (usually at a very small cost) to collect and organize the additional information. Very often these simple organizational tasks are overlooked, or the evaluator is not involved until a later stage of the project – by which time it is often too late to put the system in place.

Observing unobservables. An important application of the baseline reconstruction techniques is to identify important missing variables from the secondary data sets and to provide estimates of these variables. In the case of the previous micro-credit project, techniques such as focus groups or key informant interviews could be used to identify differences between women who did and did not apply for loans that might affect project outcomes (such as successful launch of a small business). If possible, a small sample of beneficiaries and non-beneficiaries would then be

visited to try to reconstruct information on factors such as prior business experience, attitude of the husband to his wife's business initiatives, and self-confidence. If a rapid sample survey is not possible, a second (if less satisfactory) option might be key informant interviews or focus groups providing information on project and non-project women. As always, the mixed methods approach will include quantitative methods to assess and control for selection bias. Even in strong statistical designs, good qualitative research can also help to 'observe' variables that might otherwise be a source of omitted and/or unobserved variable bias, and potentially suggest other instrumental variables that could be incorporated into regression analysis.

Identifying a range of impact evaluation design options. When real-world constraints make it impossible to use the strongest statistical designs, the evaluator should identify the range of possible design options (going from the most to the least statistically rigorous) and identify which design options are feasible within the prevailing constraints and can achieve an acceptable minimum level of methodological rigor (see below). The possible designs should be assessed in terms of their adequacy for the purposes of the present evaluation, and ways to strengthen validity through the combination of different methods should be considered.

Rapid and economical data collection methods. There are a number of techniques that can be used to reduce the costs and/or time required for data collection (Bamberger, Rugh and Mabry 2006). These include: shortening the data collection instrument by eliminating non-essential information; using more economical data collectors (school teachers or nurses instead of professional enumerators); collecting information from groups rather than individuals (focus groups etc); direct observation rather than interviews (for example to estimate community travel patterns); using secondary sources; and building required data into project monitoring and similar administrative records. All of these techniques have potential issues of quality and representativity, and as such it is important to assess these trade-offs between cost/time and validity when defining the data collection strategy.

Threat to validity. While validity should be assessed for all evaluations, the above approaches for reducing costs and time and working with less robust evaluation designs significantly increases the range and potential severity of threats to validity. The use of mixed methods also introduces additional validity issues. The use of a threats-to-validity checklist is recommended to systematically assess the potential threats, their severity and possible measures that can be taken to address them[11]. A key point of using mixed methods, however, is to triangulate data sources so as to check the validity of one instrument against another. Again, even in evaluations deploying strong statistical designs using large household surveys, qualitative methods such as anchoring vignettes (King et al 2004) can be used to ensure that survey respondents in different contexts are in fact interpreting questions in the same way.

IV. Strengthening Monitoring Systems: Understanding Project Processes and Contexts

Projects are almost never implemented exactly as planned and there are often significant variations in implementation in different project locations. Consequently, if the analysis finds no

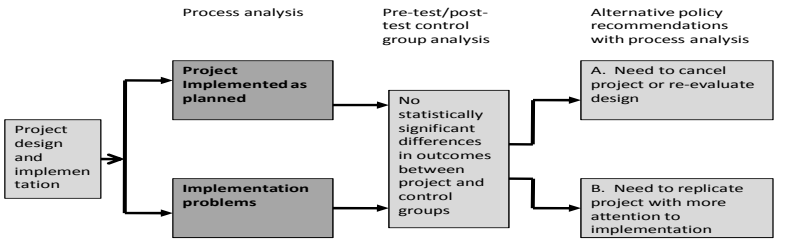
[11] A pilot version of a threats to validity checklist developed by Bamberger and Rugh (2008) is available at http://www.realworldevaluation.org/RealWorld_Evaluation_resour.html

statistically significant differences between the change in indicators of intended outcomes/impacts between the project and comparison group, the conventional evaluation designs cannot distinguish between two alternative explanations:

- The project was implemented more or less as planned, so the lack of statistical outcome differences suggests there are weaknesses in the project logic and design (at least for this particular context) and the initial recommendation would be that the project should not be replicated or at least not until it has been redesigned (alternative policy recommendation A in figure 2).
- There were significant problems or deviations during project implementation so it was not possible to test the logic of the project design. Consequently the initial recommendation might be that another pilot project should be funded with more attention to implementation (alternative policy recommendation B in figure 2).

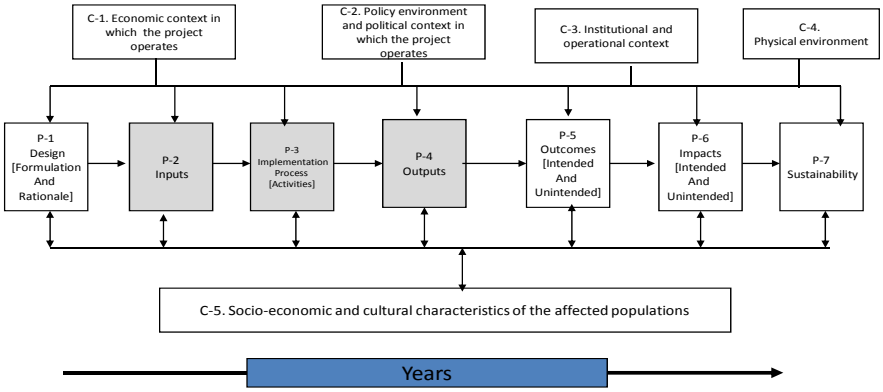
These two alternatives are shown in Figure 2.

Figure 2. Alternative policy recommendations when pretest-posttest control group comparisons find no statistically significant differences in outcome indicators



A weakness of many conventional impact evaluation designs is that they use a pre-test/posttest comparison group design that only collects data at the start and end of the project and does not examine what happened during the process of project implementation and how this might have affected outcomes. Consequently, as shown in figure 2, these designs are not able to distinguish between the alternative explanations of *design failure* and *implementation failure*. Mixed method designs, by combining statistical analysis with techniques for monitoring implementation, can test both of these hypotheses and hence considerably enhance the operational and policy utility of the evaluation.

Figure 3. Contextual and process analysis



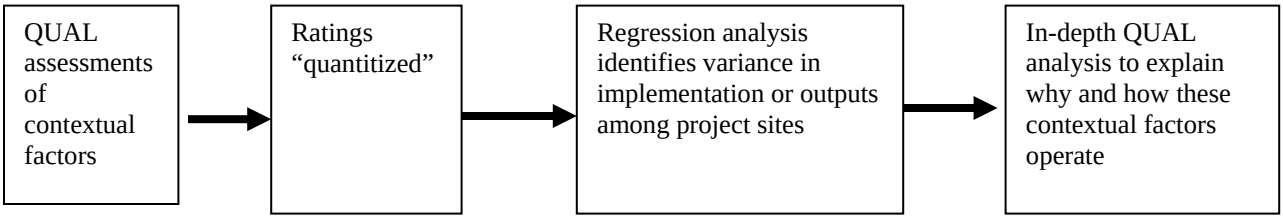
3

Figure 3 presents a stylized model that provides a framework for assessing both the quality of project implementation and the level of conformity to the original implementation plan and contextual factors that affect both overall implementation and factors explaining variations of performance in different project settings. Steps P-1 through P-7 describe the typical stages of the project implementation process, while Boxes C-1 through C-5 identify five sets of contextual factors such as the local, regional and national economy; the policy and political environment; the institutional and operational environment; the physical environment (droughts, floods, soil erosion etc); and the socio-economic and cultural characteristics of the affected populations that can affect implementation and impacts in different locations.

While some contextual factors can be measured quantitatively (for example, percentage changes in indicators such as unemployment), others will usually be measured qualitatively (for example, assessing whether the local political context is supportive of, neutral or opposed to the project[12]; the capacity and performance of local agencies involved in the project). The decision whether to use quantitative and/or qualitative indicators will also depend on data availability, so that in one country opinion polls or citizen report cards may provide detailed statistical data on attitudes to the performance of public service agencies, while in another country the rating may be based on the opinion of the local consultant or the pooled views of several key informants. For large projects with a longer time horizon and more resources, it may be possible to commission local researchers to conduct an in-depth diagnostic study covering all of these factors, but more commonly the ratings will be based on rapid data collection. Often mixed method data transformation techniques will be used so that qualitative ratings are aggregated into dummy variables that can be incorporated into regression analyses. Where time and resources permit this can be an iterative process (Figure 4) in which researchers return to the field to conduct in-depth analysis to explain how the different contextual factors influence implementation and outcomes.

[12] On attempts to assess such political contextual variables, see Barron, Diprose and Woolcock (forthcoming).

Figure 4 An iterative process for analyzing how contextual factors affect project outcomes



While contextual analysis examines the influence of external factors, process analysis looks at the internal organizational processes through which the project is implemented. However, this distinction is not rigid as external factors can affect implementation processes (for example, political pressures to provide benefits to non-eligible groups or problems within partner agencies that can affect the provision of certain services). Process analysis mainly focuses on steps P-2 (inputs), P-3 (processes) and P-4 (outputs). Some of the main sources of information for process analysis are the project's internal monitoring and administrative systems. While in theory a well-functioning monitoring system could provide most of the information required for process analysis, in practice monitoring systems often are seen as having only a narrow accountability function and often only generate a limited set of quantitative indicators. Often (though not always), the quality of the information is poor and relies on self-reporting, where the staff completing the information have an incentive to avoid problems by checking the boxes saying that progress is at least satisfactory. However, even when the quality of data is quite satisfactory, most systems rely on quantitative indicators with little attention given to the quality of the service delivery, how effectively beneficiaries are involved in planning and management, which sectors of the target population do and do not benefit, or the opinions of beneficiaries. A common example of the lack of information on beneficiary participation is the fact that many monitoring systems provide little or no information on differences in male and female participation rates or access to services, or similar information for different ethnic groups. As monitoring systems are managed by the implementing agencies, issues such as corruption, major absenteeism (by teachers, medical staff or office staff) are rarely addressed.

The mixed method approach to process analysis will build on, and try to strengthen, monitoring and other administrative reporting systems while complementing these with a variety of quantitative and qualitative techniques. These may include: participant observation (visiting schools, clinics, small business programs etc to observe how they operate); key informant interviews (including a wide range of different types of informant); focus groups; PRA and other participatory group interview techniques; selecting a panel of informants who are visited regularly throughout the life of the project; citizen report card surveys (in which a large and randomly selected sample of the target population are interviewed about their experiences with, and opinions about, public service agencies). It is also possible to make more creative use of administrative records and communications through the use of content analysis, review of e-mails and other internal communications, and analysis of the information and messages provided to the public through posters, notice-boards, newsletters and radio announcements. The information from all of these sources is used to compare how the project was actually implemented with the plan as stated in the operational manner and other planning documents.

The results of the analysis can be used for three complementary purposes. The first is to provide feedback to project management and policymakers to strengthen project implementation and to identify any groups or sectors of the target group that are not fully benefiting (and perhaps to identify non-eligible groups that are benefiting). The second purpose is to contribute to the evaluation by helping interpret the extent to which failure to achieve intended outcomes can be attributed to weaknesses in the project design or to problems during implementation. When combined with contextual analysis, this can also help explain differences in outcomes in different project settings.

The third application of the combined process and contextual analysis is to provide guidance on potential project replicability. Conventional statistical evaluation designs match project and comparison groups, which in turn provide a robust analysis of differences between *typical* project and *typical* comparison group members, but by definition makes it impossible to determine how effectively the positive project outcomes could be generalized to different situations. In other words, the policy recommendation from a conventional statistical impact design would be to say: "*If the project were replicated to a very similar population group these are the results that could be expected*". However, the analysis does not permit recommendations on how the project would be likely to perform with different population groups or in different contexts. The combination of contextual and process analysis has the capacity to provide at least some general guidelines on how the project is likely to perform in different settings and with different population groups.

V. Using Mixed Methods to Strengthen Qualitative Evaluations

A frequent criticism of qualitative methods is that they use small and often non-representative samples. Mixed methods approaches offer some strategies for responding to these concerns. For example, even if it is desirable (for whatever reason) to use small samples, researchers can sometimes make careful (and highly strategic) choices about those samples if quantitative information on the larger 'universe' of cases, and thus the nature of the distribution of key variables, is available. For example, selecting a small number of cases that are significantly different, as measured by variables collated in (say) a census or a large household survey, can give qualitative researchers a stronger basis on which to say something about the larger population. Mixed method sampling has great flexibility to select small samples with a range of different characteristics that can be used to strengthen the understanding and interpretation of quantitative data. Some of the options include selecting extreme cases, selecting cases that are typical of each major group covered by the sample, identifying cases that challenge the research hypotheses or that do not conform to the main patterns, or using reputational sampling to select people that are outstanding or particularly interesting. A well designed mixed method data collection and analysis strategy permits a series of iterations whereby statistical findings can be fed back to the qualitative researchers who then generate additional hypotheses or indicators that can be explored statistically and that then generate additional questions to be explored in the field (see Jha et al 2007).

Dynamic mixed methods designs use and analyze qualitative data quite differently than conventional approaches that only use qualitative methods for initial diagnostic studies to help develop the data collection instruments for the quantitative studies. One practical but little used

approach is to reserve a limited amount of time and resources to return to the field once the quantitative analysis has been completed and the draft evaluation report prepared. These resources are used to either provide more descriptive data on particularly interesting findings or to explore inconsistencies in the data (which often lead to some of the most interesting findings). For example, a study of village water management in Indonesia found that all except one of the village water supply systems were managed by women. In the one exception it was initially assumed that this was a reporting error and the initial reaction was to ignore it. However, the local researchers were able to return to this village and it was discovered that this was the only area in which dairy farming was practiced. In this region only women manage dairy cattle, and as this was a very profitable activity the men agreed to manage the water supply to free their wives to exploit this income generating opportunity. In none of the other villages did women have opportunities for profitable economic activities so they were assigned to manage the water supply (Brown 2000). This proved to be one of the most interesting findings of the study, but it could have easily been overlooked. How many other similarly interesting findings are never discovered because of the lack of a flexible mixed method design?

With larger sample sizes, mixed methods strategies can also take advantage of new software for analyzing qualitative data that enables reams of coded textual information to be aggregated into more manageable discrete variables. Such techniques have proved useful in analyzing text from newspapers and transcripts of interviews from village meetings (see below). The primary comparative advantage of qualitative methods, however, is their capacity to unpack the details and idiosyncrasies of local contexts and the processes by which different interaction effects play out. This does not mean that 'large N' qualitative work is a contradiction in terms; rather that the larger the 'N' from a qualitative standpoint, the greater the need and opportunity for fruitful engagement with quantitative approaches.

VI. Applying Mixed Methods in International Development Programs

This section discusses some of the special challenges in using mixed method approaches in project evaluations conducted in developing countries. While mixed method approaches are becoming more widely accepted among program evaluators in developed countries (although there is still more resistance to their use than is often claimed), there are a number of additional challenges for the application of these approaches in developing countries. Mixed method designs usually involve additional costs, time and logistical challenges (particularly when working in remote rural areas) or where there are security issues (meaning that arrival and departure times of the data collectors may have to be coordinated with local security authorities). It is also often the case that the professional divisions among different disciplines and research centers are often much greater, so that building a multi-disciplinary team can be more time-consuming and challenging. Many of the central planning and finance ministries that have a major role in approving research have a strong quantitative research tradition and may need to be convinced that the qualitative component of mixed methods is genuine "professional research". Finally, qualitative research is often associated with radical academics or civil society organizations who government agencies fear will deliberately give an anti-government slant to their evaluation findings (by, for example, using purposive sampling to select individuals or communities that are known to be critical of government programs).

We now present three brief case studies illustrating ways in which mixed method approaches have been used in developing countries.

Indonesia: The Kecamatan Development Project

The Kecamatan Development Project (KDP) in Indonesia is one of the world's largest social development projects. Implemented in the aftermath of the Suharto era and the East Asian financial crisis in 1998, KDP was primarily intended as a more efficient and effective mechanism for getting targeted small-scale development assistance to poor rural communities, but it was also envisioned as a project that could help to nurture the proto-democratic state at the local level. KDP requires villagers to submit proposals for funding to a committee of their peers, thereby establishing a new (and, by design, inclusive) community forum for decision making on development issues (Guggenheim 2006). Given the salience of conflict as a political and development issue in Indonesia, a key evaluation question is whether these forums are in fact able to complement existing local-level institutions for conflict resolution and in the process help villagers acquire a more diverse, peaceful, and effective set of civic skills for mediating local conflict. Such a question does not lend itself to an orthodox standalone quantitative or qualitative evaluation, but rather to an innovative mixed-method approach.

In this instance, the team decided to begin with qualitative work, since there was relatively little quantitative data on conflict in Indonesia and even less on the mechanisms (or local processes) by which conflict is initiated, intensified, or resolved[13]. Selecting a small number of appropriate sites from across Indonesia's 3,500 islands and 350 language groups was not an easy task, but the team decided that work should be done in two provinces that were very different (demographically and economically), in regions within those provinces that (according to local experts) demonstrated both a "high" and "low" capacity for conflict resolution, and in villages within those regions that were otherwise comparable (as determined by propensity-score matching methods) but that either did or did not participate in KDP. Such a design enabled researchers to be confident that any common themes emerging from across either the program or non-program sites was not wholly a product of idiosyncratic regional or institutional capacity factors. Thus quantitative methods were used to help select the appropriate sites for qualitative investigation, which then entailed three months of intensive fieldwork in each of the eight selected villages (two demographically different regions by two high/low capacity provinces by two program/non-program villages).

The results from the qualitative work – useful in themselves for understanding process issues and the mechanisms by which local conflicts are created and addressed (see Gibson and Woolcock 2008) – fed into the design of a new quantitative survey instrument, which will be administered to a large sample of households from the two provinces and used to test the generality of the hypotheses and propositions emerging from the qualitative work. A dataset on local conflict was also assembled from local newspapers. Together, the qualitative research (case studies of local conflict, interviews and observation), the newspaper evidence, data on conflict from national-level surveys and key informant questionnaires provided a broad range of evidence would was used to assess the veracity of (and where necessary qualify and contextualize) the general

[13] The details on this methodological strategy are provided in Barron, Diprose and Woolcock (forthcoming).

hypotheses regarding the conditions under which KDP could (and could not) be part of the problem and/or solution to local conflict.

India: Panchayat Reform[14]

A recent project evaluating the impact of "*panchayat* (village government) reform" – democratic decentralization in rural India – combines qualitative and quantitative data with a randomized trial. In 1992 the Indian government passed the 73rd amendment to the Indian constitution to give more power to democratically elected village governments (*Gram Panchayats* – henceforth GPs) by mandating that more funds be transferred to their control and that regular elections be held, with one-third of the seats in the village council reserved for women and another third for "scheduled castes and tribes" (groups who have traditionally been targets of discrimination). It was also mandated that a deliberative space – village meetings (*gram sabhas*) – be held at least two times a year to make important decisions such as the selection of beneficiaries for anti-poverty programs, and discussing village budgets.

It is widely acknowledged that the state of Kerala has been by far the most effective in implementing the 73rd amendment. There were two elements to this success. The first was that the state government devolved significant resources to the GPs with 40% of the state's expenditures allocated to them; the second element was the "People's Campaign", a grassroots training and awareness-raising effort to energize citizens to participate, with knowledge, in the panchayat system. This led to better village plans, widespread and more informed participation, and more accountable government. Kerala is, of course, a special case with very literate and politically aware citizens (literacy rates are close to 100%). The crucial policy question is whether the Kerala experiment can be replicated in much more challenging, and more representative settings.

The northern districts of the neighboring state of Karnataka represent such settings. The literacy rate is about 40%, with high levels of poverty and a feudal social environment with high land inequality. These districts are also known to be beset by corruption and extremely poor governance. If a People's Campaign could work in these districts, it could provide an important tool to transform the nature of village of democracy in the country by sharply increasing the quality and quantity of citizen participation in the panchayat system and, in turn, have a significant effect on the standard of living. Also, these districts have access to two large national schemes that have substantially increased the funding of GPs, raising the budget of GPs from about 200,000 Indian rupees a year to approximately 4,000,000 rupees. Thus GPs in these districts have fulfilled the first element of the Kerala program – high levels of funding. The evaluation focuses on assessing the impact of the People's Campaign. It randomly assigns 50 GPs as "treatment". Another set of GPs, matched to belong to the same county as the treatment GPs and with similar levels of literacy and low caste populations and randomly chosen within this subset, have been selected as "control" GPs. (They are also chosen to be at least one GP away from treatment GPs to avoid treatment spillover problems.)

[14] Other examples of mixed methods research in India include Rao (2000, 2001), Rao et al (2003) and Jha et al (2007).

The "treatment" consists, initially, of a two week program conducted by the Karnataka State Institute of Rural Development, which is responsible for all panchayat training in the state and has extensive experience in the field. The program trains citizens in participatory planning processes, deliberative decision making, and disseminates information about the programs and procedures of the panchayat. At the end of two weeks, a village meeting is held where priorities are finalized and presented to local bureaucrats. At a meeting with the bureaucrats an implementation agreement is reached wherein the bureaucrats commit to providing funding and technical support for the selected projects over the course of the year. Following this initial training, the GP is monitored with monthly two-day visits over a period of two years in order to ensure the program's progress.

An extensive quantitative baseline survey was implemented in the 200 treatment and control villages randomly selected from the 100 selected GPs, and completed a month prior to the intervention. The survey instruments, developed after several weeks of investigative field work and pre-testing, included village-level modules measuring the quality and quantity of public goods, caste and land inequality in the village, and in-depth interviews with village politicians and local officials. Twenty households from each village were also randomly chosen for a household questionnaire assessing socio-economic status, preferences for public goods, political participation, social networks and other relevant variables. Two years later the same sample of villages and households were re-interviewed with identical survey instruments. These pre-test and post-test quantitative data provide a "gold-standard" quantitative assessment of impact using a randomized trial.

To understand "process" issues, however, equal attention was given to in-depth qualitative work. A subset of five treatment and five control GPs from the quantitative sample was selected purposively for the qualitative investigation. They were selected to compare areas with low and high literacy, and different types of administrative variation. A team of qualitative investigators visited these villages for a day or two every week over a two year period investigating important dimensions of change: political and social dynamics, corruption, economic changes, and network affiliation, among other things. Under the supervision of two sociologists, the investigators wrote monthly reports assessing these dimensions of change. These reports provide a valuable in-depth look at month to month changes in the treatment and control areas that allow the assessment of the quality of the treatment, changes introduced by the treatment, and other changes that have taken place that are unrelated to the treatment. Thus, the qualitative work provides an independent qualitative evaluation of the People's Campaign, but also supplements findings of the quantitative data[15].

An important challenge in understanding the nature of 73rd amendment is to study participation in public village meetings (*gram sabhas*) held to discuss the problems faced by villagers with members of the governing committee. Increases in the quality of this form of village democracy would be a successful indicator of improvements in participation and accountability. To analyze this, a separate study was conducted on a sample of 300 randomly chosen villages across four South Indian states, including Kerala and Karnataka. Retrospective quantitative data on participation in the meetings, however, are very unreliable because people's memories are limited about what may have transpired at a meeting they may have attended. To address this

[15] This evaluation is still ongoing and results are not yet available.

issue the team decided to record and transcribe village meetings directly. This tactic provided textual information that was analyzed to observe directly changes in participation (see Rao and Sanyal 2009 and Ban and Rao 2009). Another challenge was in collecting information on inequality at the village level. Some recent work has found that sample-based measures of inequality typically have standard errors that are too high to provide reliable estimates. PRAs were therefore held with one or two groups in the village to obtain measures of land distribution within the village. This approach proved to generate excellent measures of land inequality, and since these are primarily agrarian economies, measures of land inequality should be highly correlated with income inequality. Similar methods were used to collect data on the social heterogeneity of the village. All this PRA information has been quantitatively coded, thus demonstrating that qualitative tools can be used to collect quantitative data. In this example the fundamental impact assessment design was kept intact, and both qualitative and quantitative data were combined to provide insights into different aspects of interest in the evaluation of the intervention.

Eritrea: Community Development Fund

The Eritrean Community Development Fund (CDF) was launched soon after Eritrea gained independence in the early 1990s, with two objectives: developing cost-effective models for the provision of community infrastructure (schools, health care centers, water, environmental protection, veterinary and feeder roads), and strengthening the participation of the local communities in the selection, implementation and maintenance of the projects. Separate evaluations were conducted to assess the implementation and impacts of each of the six components. This case describes how mixed methods were used to strengthen the evaluation of the feeder roads component (similar approaches were used to assess the health and education components). Three feeder roads were being constructed, each between 50-100 kilometers in length and each serving many small villages that currently had no access to roads suitable for vehicular traffic.

The evaluation was not commissioned until work had already begun on each of the three roads, but none of which was yet completed (planning and construction took on average around one year with work often interrupted during the rainy season). The evaluation had a relatively modest budget and no baseline data had been collected prior to the start of road construction. However, the CDF was intended as a pilot project to assess the efficiency and socio-economic outcomes of each of the six project components with the view to considering replication in a follow-up project. Consequently, policymakers were very interested in obtaining initial estimates, albeit only tentative, of the quantitative impacts of each component. Given the rapidly changing economic and social environment during the first decade of independence, it was recognized that the changes observed over the life of the different project components could not be assumed to be due to the project intervention. And the need for some kind of simple attribution analysis was recognized, despite the absence of a conventional comparison group.

The possibility was first considered of trying to identify areas with similar socio-economic characteristics but which did not have access to a feeder road and that could serve as a comparison group. However, it was concluded, as is often the case with the evaluation of the social and economic impact of roads, that it would be methodologically difficult to identify

comparable areas and in any case extremely expensive to conduct interviews in these areas, even if they could be found. Consequently the evaluation used a mixed-method design that combined a number of different data sources and that used triangulation to assess the validity and consistency of information obtained from different sources. The evaluation combined the following elements:

- The evaluation was based on a program theory model that described the steps and processes through which the project was expected to achieve its economic and social impacts and that identified contextual factors that might affect implementation and outcomes. The theory model also strengthened construct validity by explaining more fully the wide range of changes that road construction was expected to achieve so that impacts could be assessed on a set of quantitative and qualitative indicators. Some of the unanticipated outcomes that were identified in this way included: strengthened social relations among relatives and friends living in areas that were previously difficult to reach, and strengthened and widened informal support networks as people were able to draw on financial, in-kind and other support from a geographically broader network.
- Quantitative survey data was obtained from a stratified sample of households along the road who were interviewed three times during and after the road construction (the evaluation started too late for a pre-test measure)
- The baseline conditions of the project population prior to road construction were "reconstructed" by combining: recall of the time and cost for travel to school, to reach a health center, to transport produce to markets, and to visit government agencies in the nearest towns; with information from key informants (teachers, health workers, community leaders etc); and data from secondary sources. Estimates from different sources were triangulated to test for consistency and to strengthen the reliability of the estimates.
- Data on comparison groups, before, during and after road construction were obtained from a number of secondary sources. Information on school on attendance by sex and age were obtained from the records of a sample of local schools. In some cases the data also included the villages from which children came so that it was possible to compare this information with recall from the interviews in project villages. Records from local health clinics were obtained on the number of patients attended and the medical services provided. Unfortunately the records did not permit an analysis of the frequency of visits of individual patients so it was not possible to estimate whether there were a relatively small number of patients making frequent use of the clinics or a much larger number making occasional visits. Most of the local agricultural markets were cooperatives that kept records on the volume of sales (by type of produce and price) for each village so this provided a valuable comparison group. It was planned to use vehicle registration records to estimate the increase in the number and types of vehicles before and after road construction. However, qualitative observations revealed that many drivers "forgot" to register their vehicles so this source was not very useful.
- Process analysis was used to document the changes that occurred as road construction progressed. This combined periodic observation of the number of small businesses along the road, changes in the numbers of people travelling and the proportions on foot, using animal traction, bicycles and different kinds of vehicles.

- Country-level data on agricultural production and prices, available over a number of years, provided a broader picture and also to correct for seasonal variations in temperature and rainfall (both between different regions and over time). This was important in order to avoid the error of measuring trends from only two points in time.

All of the data sources were combined to develop relatively robust estimates of a set of social and economic changes in the project areas over the life of the project, and to compare these changes with a counterfactual (what would have been the condition of the project areas absent the project) constructed through combining data from a number of secondary sources. The credibility of the estimates of changes that could be (at least partially) attributed to the project intervention was then tested through focus groups with project participants, discussions with key informants, and direct observation of the changes that occurred during project implementation.

This evaluation, conducted with a relatively modest budget, and drawing on the kinds of secondary data and recall information that are often available, illustrates how mixed method designs can offer a promising approach to developing an alternative to the conventional statistical counterfactual, thus strengthening our understanding of the potential impacts of the majority of projects where the conventional counterfactual cannot be applied.

VII. Conclusion: Continuing Challenges and Opportunities

Challenges

While dramatic progress has been made in the application of mixed methods in the US and some other industrial nations, and despite a slow but steady increase in published studies in developing countries, a number of challenges continue to face evaluators wishing to apply mixed methods in the monitoring and evaluation of development projects and policies in these contexts.

A first challenge is the fact that mixed methods have been the evaluation design of choice for many development agencies for many years. However, many of these evaluations used somewhat ad hoc approaches and most do not apply the kinds of methodological and conceptual rigor that is required by academic journals such as the Journal of Mixed Method Research. So the mixed method approach is not new per se, but the professional, financial and other resources have usually not been available to increase methodological rigor.

While it is claimed (although not everyone would agree) that the “paradigm wars” are long ended in the US and Europe and that there is a general acceptance of mixed method approaches, in many, but certainly not all, developing countries the divisions between quantitative and qualitative researchers are still quite pronounced. In some countries qualitative researchers tend to be (or are perceived as being) more politically radical than their quantitative colleagues, so this can provide a further complication. Indeed, for precisely this reason. autocratic regimes are often amenable to training engineers and statisticians to help ‘manage’ the state but are much less sympathetic towards disciplines such as anthropology, sociology and journalism that might unearth material questioning the claims and legitimacy of the regime. Even in more open societies, this often means that, in practical terms, considerable time and effort is usually required to identify and build a team of local researchers who can work well together on a mixed

methods approach. Unfortunately the planning of many development evaluations does not allow much time for team building, so often the result can be two separate strands of quantitative surveys not very clearly linked to in-depth case studies or other qualitative data collection.

A related challenge in many developing countries has been the fact that most senior officials in central finance and planning agencies have been trained in economics and quantitative methods and at least until recently many have been suspicious of qualitative methods that are "not really professional research" so many research and consulting agencies have had little incentive to develop a mixed method capacity. In the last few years many research agencies are trying to strengthen their capacity in mixed method approaches but this continues to be a weak area.

Another practical problem is the lack of local expertise in mixed method research in many countries. The more rigid division between quantitatively and qualitatively-oriented university faculties also means that in many countries not many courses are offered on mixed methods. It has been our experience that many university-based consulting centers that have a strong track record in quantitative research have found it difficult to integrate a solid qualitative component to their evaluation. Too often a few somewhat ad hoc focus groups are tacked on to the end of the quantitative evaluation with no clear link to the rest of the study.

There are also many practical logistical challenges in applying mixed methods in many development contexts. Many development evaluations require data collection in remote rural areas or in urban areas where security threats makes it more difficult to use the flexible approaches required by mixed methods. For example, in some studies all data collectors have to be transported together to remote locations and all need to arrive and leave at the same time so that qualitative interviewers do not have the flexibility to take advantage of invitations to social events such as weddings, parties, funerals that are excellent opportunities to observe the community in action. In many countries police or military authorities may also require exact details on who is to be interviewed.

Opportunities

The growing interest in mixed methods, combined with recognition of the special challenges, presents a number of exciting opportunities for strengthening the use of mixed methods in development evaluations.

A first area of opportunity is in developing effective monitoring systems. While there is a widespread acceptance that monitoring is essential for good project implementation, much less thought has been given to the design of monitoring systems than to methods to evaluate project impact. Mixed methods with their ability for the rapid and economical collection of data customized to the characteristics of a particular program and providing information that can be easily contextualized can be very helpful in monitoring design.

While well designed monitoring systems can potentially provide a wealth of data to strengthen the evaluation design, as we saw earlier these benefits are only achieved if monitoring and evaluation are treated as two components of a program management system. This requires that the data requirements for the future impact evaluation are built into the monitoring systems in a

way that ensures that high quality data is collected and organized in the format required for the evaluation analysis. While this may seem a relatively straightforward task, in practice it presents a major organizational challenge as the evaluation team is not normally consulted during the design of the monitoring system. During the hectic period when projects are being negotiated, staff hired and implementation arrangements put in place, project managers are often unwilling to spend extra time (and resources) to worry about the requirements of an evaluation that will not produce findings for several years. So the incentive system for management may need to be changed as well as broadening the role of the evaluation team and ensuring their involvement at an earlier stage in the project design.

A second area of opportunity is the fact that strong statistical evaluation designs, even if they are believed to be the best option, can only be applied in a small number of development evaluations. Until this situation changes, the body of literature on how to strengthen evaluation designs for the remaining (i.e., the majority of) evaluations is very limited and this provides a unique opportunity for developing mixed method designs that make the most effective use of the many, but often incomplete, sources of quantitative and qualitative information that can be generated. Mixed method designs are intended to address exactly this kind of challenge and there are strong opportunities for the application of practical, cost effective mixed method evaluations.

A third major area of opportunity concerns developing alternatives to the conventional statistical counterfactual. There is now widespread acceptance that the use of randomized trials, and statistically matched project and control designs, can provide a valid counterfactual. More generally, there is an increasing awareness that all development agencies need to be able to address the question: “How do we know that the observed changes in the project population can be attributed to our intervention?”, or put another way: “How can we estimate what would have been the condition of the project population if our project had not taken place?” We have argued here that techniques such as concept mapping or qualitative group consultation methodologies such as PRA can be used to construct a counterfactual, but this is still a nascent area of research and no widely accepted alternatives to the conventional counterfactual have yet been found. This is another area of opportunity for mixed methods.

Some of the other areas of opportunity include: refining approaches to mixed method sampling; incorporating process and contextual analysis into strong statistical designs; reconstructing baseline data and using mixed methods to generate more reliable estimates of unobservables in econometric analysis; and improving the extent to which project efficacy is assessed not only with respect to a counterfactual but to where it should be relative to other projects of a particular type at a given time interval. Furthermore, the development community is moving away from support for individual projects to funding of complex multi-component, multi-agency national level development programs. There is a growing demand for new evaluation methods to assess the outcomes of these complex programs, and these are areas in which mixed methods, with their flexibility to draw on and integrate many different sources of information, can make a valuable contribution.

A final challenge for mixed method evaluation is to provide guidelines on minimum levels of acceptable methodological rigor when drawing on diverse data sources that are often collected under tight budget and time constraints or where much of the data is collected under difficult

circumstances. Conventional threats to validity analysis have still not been widely applied to mixed method evaluations, and there is an opportunity to develop standards that ensure a satisfactory level of methodological rigor while also being realistic. There is a challenge of on the one hand avoiding the "anything goes" approach of some evaluators who believe that given the difficult situations in which they are working they cannot be held accountable to normal standards of rigor, while on the other avoiding the claim of some academic researchers who would apply exactly the same criteria to the assessment of an evaluation of a program to prevent sexual harassment of women in refugee camps as they would to a large federally funded evaluation of education-to-work programs in a US city. At the end of the day, it should only be logical that knowledge claims regarding the efficacy of the diverse range of development projects, in all the diverse contexts in which they are deployed, should be made on the basis of an appropriate corresponding diversity of social science methods and tools.

References

Bamberger, Michael (2009a) "Strengthening Impact Evaluation Designs through the Reconstruction of Baseline Data" *Journal of Development Effectiveness* 1(1): 37-59

Bamberger, Michael (2009b) "Why do so many evaluations have a positive bias?" Paper presented at the Australasian Evaluation Society Annual Conference. September 3, 2009. Canberra, Australia.

Bamberger, Michael, Keith Mackay and Eileen Ooi (2004) *Influential Evaluations: Evaluations that Improved Performance and Impacts of Development Programs*. Independent Evaluation Group. Washington, DC: World Bank

Bamberger, Michael, Keith Mackay and Eileen Ooi (2005) *Influential Evaluations: Detailed Case Studies.* Independent Evaluation Group. Washington D.C. The World Bank.

Bamberger, Michael, Jim Rugh and Linda Mabry (2006) *RealWorld Evaluation: Working under Budget, Time, Data and Political Constraints* Thousand Oaks, CA: Sage Publications

Ban, Radu and Vijayendra Rao (2009) "Is Deliberation Equitable? Evidence from Transcripts of Village Meetings in South India" Policy Research Working Paper No. 4928, Washington DC: World Bank

Banerjee, Abhijit (2007) *Making Aid Work* Cambridge, MA: MIT Press

Barron, Patrick, Rachael Diprose and Michael Woolcock (forthcoming) *Contesting Development: Participatory Projects and Local Conflict Dynamics in Indonesia* New Haven: Yale University Press

Brown, Gillian (2000) “Evaluating the Impact of Water Supply Projects in Indonesia”, in Michael Bamberger (ed.) *Integrating Quantitative and Qualitative Research in Development Projects*. Washington, DC: World Bank, pp. 107-113

Cartwright, Nancy (2007) “Are RCTs the Gold Standard?” *BioSocieties* 2: 11-20

Deaton, Angus (2009) “Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development” Working Paper No. 14690, National Bureau of Economic Research, Cambridge, Massachusetts

Duflo**,** Esther and Michael Kremer (2005) “Use of Randomization in the Evaluation of Development Effectiveness” in George Pitman, Osvaldo Feinstein and Gregory Ingram, (eds.) *Evaluating Development Effectiveness* New Brunwick: Transaction Publishers, pp. 205-31

Gibson, Christopher and Michael Woolcock (2008) “Empowerment, Deliberative Development and Local Level Politics in Indonesia: Participatory Projects as a Source of Countervailing Power” *Studies in Comparative International Development* 43(2): 151-180

Guggenheim, Scott E. (2006) “Crises and Contradictions: Explaining a Community Development Project in Indonesia”, in Anthony Bebbington, Scott E. Guggenheim, Elisabeth Olson, and Michael Woolcock (eds.) *The Search for Empowerment: Social Capital as Idea and Practice at the World Bank.* Bloomfield, CT: Kumarian Press, pp. 111-144

Jha, Saumitra, Vijayendra Rao and Michael Woolcock (2007) “Governance in the Gullies: Democratic Responsiveness and Community Leadership in Delhi’s Slums” *World Development* 35(2): 230-46

Kane, Mary and William Trochim (2007) *Concept Mapping for Planning and Evaluation*. Thousand Oaks, CA: Sage Publications

King Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon (2004). “Enhancing the Validity and Cross-Cultural Comparability of Survey Research.” *American Political Science Review* 98(1): 191-207

Morgan, Stephen and Christopher Winship (2007) *Counterfactuals and Causal Inference: Methods and Principles for Social Research* New York: Cambridge University Press

Morra, Linda and Ray Rist (2009) *The Road to Results: Designing and Conducting Effective Development Evaluations* Washington, DC: World Bank

Mosse, David (2005) *Cultivating Development: An Ethnography of Aid Policy and Practice* London: Pluto Press

NONIE (Network of Networks on Impact Evaluation) (2008) *Impact Evaluations and Development. NONIE Guidance on Impact Evaluation.* Draft prepared for the Cairo International Evaluation Conference, April 2009

Olken, Ben (2007) “Monitoring Corruption: Evidence from a Field Experiment in Indonesia” *Journal of Political Economy* 115(2): 200-249

Patton, Michael Quinn (2008) *Utilization-Focused Evaluation* (Fourth Edition) Thousand Oaks, CA: Sage Publications

Pritchett, Lant (2002) “It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation” *Policy Reform* 5(4): 251-269

Rao, Vijayendra (2000) “Price Heterogeneity and Real Inequality: A Case-Study of Poverty and Prices in Rural South India” *Review of Income and Wealth* 46(2): 201–12

Rao, Vijayendra (2001) “Celebrations as Social Investments: Festival Expenditures, Unit Price Variation and Social Status in Rural India” *Journal of Development Studies* 37(1): 71–97

Rao, Vijayendra, Indrani Gupta, Michael Lokshin, and Smarajit Jana (2003) “Sex Workers and the Cost of Safe Sex: The Compensating Differential for Condom Use in Calcutta” *Journal of Development Economics* 71(2): 585-603

Rao, Vijayendra and Michael Woolcock (2003) “Integrating Qualitative and Quantitative Approaches in Program Evaluation”, in Francois J. Bourguignon and Luiz Pereira da Silva (eds.) *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools* New York: Oxford University Press, pp. 165-90

Ravallion, Martin, (2008) “Evaluating Anti-Poverty Programs,” in Paul Schultz and John Strauss (eds.) *Handbook of Development Economics* Volume 4, Amsterdam: North-Holland

Vijayendra Rao and Paromita Sanyal (2009) “Dignity through Discourse: Poverty and the Culture of Deliberation in Indian Village Democracies” Policy Research Working Paper No. 4924, Washington, DC: World Bank

Woolcock, Michael (2009) ‘Toward a Plurality of Methods in Project Evaluation: A Contextualized Approach to Understanding Impact Trajectories and Efficacy’ *Journal of Development Effectiveness* 1(1): 1-14