USAID
FROM THE AMERICAN PEOPLE

2009 DEMOCRATIC DECENTRALIZATION PROGRAMMING HANDBOOK - EXCERPT CHAPTER 6 MONITORING & EVALUATING PROGRAM IMPACT

JUNE 2009

This publication was produced for the United States Agency for International Development by ARD, Inc.

DISCLAIMER
The views expressed in this guide do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

6.0 MONITORING AND EVALUATING PROGRAM IMPACT

Having described programming strategies and activities in the three main arenas of decentralization, this chapter turns to the question: how can USAID determine whether these strategies and activities are producing the intended results? This chapter applies recent thinking on monitoring and evaluation (M&E) to the approach to decentralization advocated in this handbook.

A distinction should be drawn at the outset between two different kinds of M&E activities. The first seeks to assess progress in implementing decentralization reforms. To this end, one might gather and analyze data on what are sometimes called output indicators: the number of meetings and workshops held, officials trained, and so on. Output indicators can help to document whether necessary steps are being taken toward effective support of decentralization programs, and they may be especially useful as management tools for program implementation.

The second kind of M&E seeks to assess the impact of decentralization programming on the broader goals described in this handbook: enhancing stability, promoting democracy, and fostering economic development. The key questions here are whether and how we can attribute outcomes along these dimensions to specific USAID initiatives in support of decentralization programming. This kind of M&E is crucial, for it is the only way to assess what works and what does not in decentralization programming. Rigorous, repeated assessments of the impact of different programs provide a valuable way of making decentralization programming more effective.

This chapter focuses on this second kind of M&E, discussing how program evaluation activities can be structured to investigate the impact of USAID interventions on the broader objectives of decentralization reforms. Given the extensive and evolving USAID documentation of M&E systems geared toward program implementation, material which is already available to field officers and program implementers, this chapter will not attempt to duplicate this topic. Instead, it will concentrate on newer approaches and rigorous evaluation methodologies—including randomized evaluations—which are particularly relevant to assessing the impact of decentralization programming.

Note that there is a subtle difference between evaluating the impact of decentralization initiatives and the impact of USAID interventions in support of decentralization initiatives, though the two are intimately linked. Assessing the impact of USAID interventions in support of decentralization may tell us much about the broader impact of decentralization on outcomes such as stability, democracy, and economic development. In addition, USAID can sometimes select the subnational units with which it will work, which in principle allows for rigorous evaluation of the impact of its initiatives. This chapter therefore focuses on the impact of USAID interventions in support of decentralization initiatives.

The chapter discusses various issues that arise in measuring strategic objectives and formulating outcome indicators in decentralization programs. It reviews some of the particular opportunities and challenges of designing M&E for assessing program impact in the decentralization context. Additionally, it shows how M&E can be structured to follow the analytical principles outlined in Chapters 4 and 5.

6.1 OBJECTIVES AND OUTCOMES

Before describing methodologies for assessing the extent to which decentralization programming enhances stability, promotes democracy, or fosters economic development, it is crucial to discuss how these broad objectives may be measured. This section discusses considerations that may arise in formulating strategic objectives that relate to each of these three broad goals and in defining outcome indicators for each objective.

6.1.1 Defining Objectives and Outcome Indicators

Table 6.1 builds on Table 5.1, which identified broad strategies for promoting stability, democracy, and economic development in the national, subnational, and civil society arenas. It is crucial to disaggregate these broad concepts into strategic objectives, such as reducing conflict or increasing the responsiveness of subnational governments. The top of each cell of Table 6.1 lists key objectives for each concept and in each arena. These objectives are proposed with the four key dimensions of decentralization (authority, autonomy, accountability, and capacity) in mind. Clearly, some objectives relate more to one dimension than another.

The bullet points beneath each objective in the table suggest potential outcome indicators for that objective. For instance, an outcome indicator relevant to the goal of reducing conflict and contentious action in the national arena might be the number of marches or protests nationwide during a given time period. An outcome indicator related to the objective of increasing responsiveness of subnational governments to citizens might be the proportion of citizens who positively evaluate the responsiveness of municipal or provincial government to their needs and demands. An indicator relevant to the objective of increasing the ability of CSOs to partner with subnational governments and administrations might include the number of public-private partnerships for infrastructure projects or the extent of private investment in employment-generating projects.

In practice, specific measurement strategies would be required to gather data on each of the outcome indicators listed in the bullet points in Table 6.1. The unit of measure must be identified, data sources found or created, the frequency of data collection defined, and the party responsible for data collection and reporting pinpointed. The specific strategies will vary widely. For example, the number of marches or protests in a given jurisdiction and time period might be culled by local partners from reports of local human rights organizations or other CSOs, local newspapers, or other sources. The proportion of citizens who evaluate local government responsiveness positively might be estimated by hiring a local survey firm to survey citizens in relevant jurisdictions before, during, and after program implementation. The extent of private investment or CSO involvement in public sector projects might be measured through surveys of local businesses and CSOs, among other strategies. The sources, frequency of data collection, and responsible party will differ as a function of the strategic objective being pursued, the evaluation issues discussed in Section 6.2 below, and other factors.

Thus, there will not be a single set of outcome indicators or measures that will be useful in all contexts. Like programming strategies themselves, the conceptualization and measurement of appropriate outcomes cannot be approached through a standardized prescription. Complexities and context must be taken into account. Specific program monitoring plans may call for different kinds of outcome indicators, depending on the strategic objectives of the program.

Table 6.1, therefore, provides examples—for each of the three primary goals of decentralization—of one or two objectives and some ways of formulating outcome indicators for them. In practice, of course, many more objectives may be relevant. In addition, some of the outcomes may be easier to measure than others. For example, while it may be relatively straightforward to measure deaths per capita in violent conflicts, it is undoubtedly more difficult to estimate the extent to which standards are obeyed in the national arena.

The outcomes in Table 6.1 are not necessarily shaped by USAID interventions, but they may well be. After all, the goal of evaluation is to find out whether decentralization programming impacts these objectives and outcomes—and what kind of programming works best. For example, expanding political participation in subnational units or strengthening the ability of deconcentrated national ministries to provide desired goods and services might reduce marches and protests per capita or deaths in violent conflicts per capita. Promoting participatory budgeting at the local level might increase local government responsiveness to citizen demands and therefore raise the proportion of citizens who rate local government responsiveness highly. Matching the desired outcome to the intervention in question is key to formulating strategic objectives and designing appropriate evaluations.

It should also be borne in mind, as discussed in Section 6.2 below, that observing changes in outcome indicators over the life of a project does not on its own serve as a reliable basis for inferring program impact. Yet before we can assess the causal impact of USAID programming, it is important to formulate indicators for the desired outcomes. Formulating a chart of strategic objectives and outcome indicators like Table 6.1 and then designing specific measurement strategies is crucial to successful M&E for assessing program impact.

TABLE 6.1. FORMULATING OBJECTIVES AND OUTCOME INDICATORS: EXAMPLES

OBJECTIVES	**National Arena**	**Subnational Arena**	**Civil Society**
Stability	***Reduced conflict and contentious action*** • Number of marches or protests per capita nationwide • Number of deaths per capita in violent conflicts nationwide ***Increased coordination among national agencies responsible for decentralization*** • Frequency of each agency's administrative contact with other agencies ***Improved national policy environment***	***Reduced conflict and contentious action*** • Number of marches or protests per capita in subnational units • Number of deaths per capita in violent conflicts in subnational units ***Increased capacity of local administrative units engaged in service delivery*** • Number of staff • Degree to which bureaucratic recruitment is based on merit ***Heightened coordination among subnational units*** • Frequency of local	***Increased organization across ethnic lines inside CSOs*** • Ethnic heterogeneity among CSO members ***Improved coordination between traditional authority structures and subnational entities*** • Frequency of local authorities' contact with national authorities ***Increased capacity of CSOs*** • Size of CSOs' operating budgets ***Decreased dependence***

OBJECTIVES	National Arena	Subnational Arena	Civil Society
	regarding decentralization • Existence of comprehensive decentralization legislation	authorities' contact with authorities in other localities	***of CSOs on external revenue sources*** • Degree of concentration of CSO revenue sources
Democracy	***Increased competitiveness and inclusiveness of national elections*** • Degree to which political parties may compete freely for national office. • Ease with which political parties can register to participate in elections • Ability of independent candidates to run for office ***Improved respect for citizen rights*** • Number of human rights violations, as tracked by CSOs or ombudsman's office	***Increased responsiveness of subnational governments to citizen needs and demands*** • Proportion of citizens who positively evaluate government responsiveness to their demands ***Heightened political competition*** • Existence of competitive local elections ***Improved performance of subnational representative bodies*** • Number of mayoral initiatives opposed or vetoed by subnational councils	***Increased formal interaction between government and civil society*** • Number of CSOs participating in participatory planning and budgeting • Number of formal demands made to local government units for local policy reforms and/or improved services by CSOs • Percentage of total subnational budget under the control of participatory bodies
Development	***Increased capacities of national agencies to set and enforce regulations and standards*** • Number of administrative and technical regulations and standards publicized by national	***Improved capacities of subnational units to provide key infrastructure and services*** • Number of concrete plans for service extension or quality improvement prepared and implemented	***Increased ability of CSOs to partner with subnational governments and administrations*** • Number of public-private partnerships for infrastructure projects and service delivery

OBJECTIVES	National Arena	Subnational Arena	Civil Society
	agencies • Extent to which national government can penalize subnational units for noncompliance with standards ***Increased coordination between national agencies*** • Existence (or budget) of national-level coordinating body ***Increased clarity in basic framework*** • Revenue sharing percentage is set by law or inserted into constitution	• Extent to which recruitment of bureaucratic staff is meritocratic • Extent to which bureaucrats have long-term career incentives	• Extent of private investment or CSO involvement in public-sector projects • Rate of immigration to subnational units offering business friendly policies and employment opportunities • Percentage of subnational units that use CSOs to produce decentralized goods and services
Note: Italicized bold text indicates strategic objectives; bullets indicate outcome indicators			

6.2 INFERRING CAUSAL IMPACTS

Inferences about the causal impact of programming interventions generally must be derived from comparisons between units subjected to an intervention (treated units) and units not subjected to the same (control units). For example, when we want to know whether a participatory budgeting program at the local level influenced citizen perceptions of local government responsiveness (see Section 5.2.2), one approach would be to compare citizen perceptions before a project begins with perceptions after a project is completed, perhaps by administering a baseline survey. Alternatively, after program completion, surveys might be taken in subnational units subjected to the intervention and in those that were not, allowing for a comparison of outcomes. Sometimes, these strategies can be combined, as when one gathers data in both the treatment group and the control group before and after the programming intervention.

Each of these alternatives holds the potential for an invalid attribution of impact. Comparisons over time can be misleading, since differences before and after a program might simply reflect preexisting trends or might be due to factors other than the intervention. Suppose that surveys of citizens suggest a decrease in citizens' positive evaluations of local government responsiveness over the life of a program. Rather than being a result of the program's negative impact, this decrease may be due to other factors, such as an economic recession or a new national policy that impacted negatively on a local government's ability to respond to citizen demands. Thus, the gathering of baseline data on units at the start of a program, while important, is usually not sufficient to provide convincing evidence of program impact.

Comparisons across treated and control units after a program's completion can also be misleading. One big threat to valid inference is that treated and control units may differ in ways beyond whether the intervention was applied, and these differences may influence the outcome of interest. For instance, if USAID chose to work in poorer municipalities that had less responsive governments to begin with, citizens in the treated municipalities might rate their governments as less responsive than citizens in untreated municipalities after the program. Yet this would not be a sound reason to conclude that the intervention did not work to promote local government responsiveness, since the difference in outcomes could be attributed to preexisting factors.

Another possibility would be to compare the change in outcomes in the treated group to the change in outcomes in the control group. This offers a more secure basis for drawing inferences about program impact than simple comparisons across time or space. However, this approach can be misleading if factors that affect local government responsiveness vary across the treatment and control groups during the life of a program. For example, USAID may work in poorer rural regions, and a drought during the program may disproportionately affect subnational units in these regions, causing a negative change in citizen evaluations of local government effectiveness in the treated units.

Clearly, selecting appropriate control units for comparison to the treated units is fundamental. Ideally, in order to understand the effects of the intervention, control units should closely resemble the treated units in most other factors. Strategies for selecting control units are discussed further below.

Sometimes it will be feasible to make random choices of the units with which USAID and its partners will work. Randomized evaluation constitutes the gold standard for drawing inferences about the causal impact of programming interventions, and randomized trials are expected be increasingly used in M&E of democracy promotion activities at USAID. In addition, decentralization programs are often particularly amenable to randomized trials, especially in supporting activities in the subnational arena that involve large numbers of subnational units (Section 5.2). The next subsection introduces key ideas about randomized evaluations and their applications to decentralization programming.

6.2.1 Randomized Trials

Randomized trials roughly balance across the treated and control groups those factors—other than the intervention itself—that might influence outcomes. As a result, randomized trials provide the most secure basis for drawing inferences about the impact of programming interventions.

Suppose that a region in which USAID and its partners work contains richer and poorer subnational units, and that the richer units tend to have more responsive local governments. Randomly assigning units to treatment and control implies that, on average, an equal proportion of units in the treated and control groups will be richer. If citizens rate local government responsiveness more positively in the treated group than in the control group after the intervention (as long as this difference is too big to have reasonably occurred by chance), the difference is unlikely to be due to a variation in the wealth of units between the treated and control groups. We therefore have strong evidence that the intervention has had a causal impact.

Although the simplest randomized design assigns some units to treatment and others to control at the start of the program, there are many other options. For example, it is sometimes useful to randomize the rollout of programs to different units, particularly with interventions that are expected to have a relatively rapid effect. With such a randomized phase-in of a program, outcomes can be compared between those units treated earlier and those treated later. Many other modifications may be suitable for different contexts, questions, and types of interventions.

When is it feasible to assign units randomly for interventions? This will depend on the nature of the intervention or activity as well as on the arena in which it takes place. For instance, activities aimed at enhancing decentralization and democratic governance in the national arena may be less amenable to random assignment, because USAID and its partners often work with a single national government or legislature to pass a law or improve the policy environment for decentralization (Section 5.1).

Random assignment is more feasible when activities or interventions are aimed at the subnational arena or civil society, because there may be many subnational units or CSOs with which USAID and its partners could work. As discussed in the previous chapter, USAID will rarely have the resources necessary to conduct programming simultaneously in all subnational units in a country, so explicit choices must be made about where and how to work (Section 5.2.1).

In this context, random assignment is both feasible and attractive. Randomization offers the best basis for assessing program impact after the program's conclusion, but there may be other advantages as well. For instance, using a lottery to decide on the allocation of programming to subnational units can be seen as the fairest way to make choices about where to work.

Programs in which interventions or sets of interventions are implemented in some subnational units but not others can offer opportunities for randomized evaluations. By way of illustration, consider the decentralization program in Senegal, where the government passed a decentralization reform in 1996 that transferred new responsibilities to elected councils in 67 communes and 320 communautés rurales (Section 5.3.3). Beginning in 2000, USAID implemented a five-year program of technical assistance and training to 50 of these Senegalese local government units. In principle, choosing these 50 units at random from among 387 possible units would have offered the best means of assessing the impact of the program on budget management capacities and revenue mobilization. Measures of outcome indicators could then be compared across the treated and control units, and average differences could be attributed to participation in the program.

Note that choosing the units at random does not necessarily imply what the intervention in treated units will actually be. In Senegal, for example, though treated units were subject to a standard set of interventions such as the budget forum and basic training in roles and responsibilities, a significant feature of the implementation strategy was demand driven. Participants would articulate their most important needs for public goods and services and then rank them by priority during town hall meetings. Even if the nature of the intervention is shaped by participating units in this fashion, the random assignment of units to treatment and control still allows an assessment of the causal impact of participation in the USAID-sponsored program. It should also be noted that, in a country like Senegal, the cost of an intervention will depend upon the accessibility of the location. As discussed further below, from the perspective of program evaluation, this issue can be addressed by identifying beforehand those locations that pose an acceptable level of cost and then randomly selecting units for treatment and control from within this list.

As another example, USAID/Peru launched a program in 2002 to support national decentralization policies initiated by the Peruvian government. Over a five-year period, the program was intended to support the implementation of mechanisms for citizen participation with subnational governments (such as participatory budgeting), to strengthen the management skills of subnational governments in selected regions, and to increase the capacity of nongovernmental organizations in these same regions to interact with their local governments. Interventions took place in more than 500 subnational units at the municipal and departmental levels. Though in this program (as in Senegal) treated units were not selected at random, it would have been feasible to do so. The decentralization program in Peru focused on municipalities in seven non-randomly selected regions in which USAID has historically worked. The point here is that municipalities could have been chosen randomly for treatment and control from within this set of seven regions.

Besides offering the best way to evaluate whether a program or intervention worked to achieve desired impacts, randomized trials also offer a way to assess which interventions work best. In principle, it is possible to assign some subnational units to one set of treatments and other subnational units to a different set. If the assignment is random, outcomes can be compared across the two groups (and potentially to a control group), and the relative efficacy of the different interventions can be measured. For example, one might like to know whether local government effectiveness is best promoted in a particular decentralization program by targeting elected councils or civil society organizations. Randomized designs offer a rigorous way to answer to this question.

6.2.2 Objections to Randomized Trials

Are randomized evaluations of interventions associated with decentralization programming really feasible? Several objections may arise in the minds of USAID field officers and local partners. We discuss here what those objections may be and suggest how randomized designs can be modified or tailored to accommodate these objections.

Political will. Perhaps the most common objection to randomized designs concerns the importance of political will; that is, the need to obtain the consent and cooperation of local authorities and other actors who may be key to the success of decentralization programming. This is clearly an important issue since, without local cooperation, USAID cannot be effective in its support strategies. This leads to the claim that it is preferable to "build on the best" (and neglect "the rest") by choosing locations that present the most receptive programming environments.

While this may be a reasonable strategy for assuring project success, comparing outcomes in those subnational units that agree to work with USAID to those that do not can result in misleading assessments of program impacts. Locations that present the most receptive programming environments may have more responsive and accountable governments from the start. This will undermine inferences about the impact of the decentralization programming on government responsiveness when those inferences are based on comparisons across treated and untreated units after a program's conclusion.

There are several potential solutions to this problem. One solution is to create a list of units that would be willing to work with USAID and to assign units randomly to treatment and control from within this list. Another possibility is based on the intention-to-treat principle in experimental analysis. Here randomization to treatment and control can occur within the group of all eligible units, not just those who would be most willing to work with USAID and its partners. Suppose there are 500 subnational units with which USAID could work in a participatory budgeting program. Some 250 of these might be randomly invited to participate in the program, with other locations serving as controls. In this way, both the treated group and the control group would contain some of the "best" and some of the "rest."

The intention-to-treat principle treats non-cooperating subnational units like those patients assigned to treatment in a medical trial who do not take the pill. With 250 units in treatment and 250 units in control, however, the numbers of noncomplying units will be roughly balanced across the treatment and control groups. Comparing government responsiveness across the treatment and control groups can provide a valid way of estimating the effect of the participatory budgeting program. This intention-to-treat principle is illustrated in the text box below in the context of a hypothetical experimental design for studying the effects of a rollout of telecenters to subnational units.

Use of the intention-to-treat principle carries some implications for program design. If it is anticipated that some units will refuse to participate, it may be useful for USAID staff and partners to estimate the refusal rate and adjust the number of randomized invitations accordingly. Another practical issue involves securing the consent of the government and other actors to the randomization of invitations. In some programs, this is not an important obstacle, once a list of eligible municipalities is created. Since only some units on the list will be treated, the only issue is how to choose the units for treatment, and a lottery may often be perceived as a fair way to do so. Related issues and solutions are discussed next.

Decentralization and Rural Connectivity: The Intention-To-Treat Principle

In 2007, the Government of Peru approached USAID/Peru for assistance with the rollout of community computer centers (also known as telecenters) in selected rural municipalities. The plan called for USAID to fund initial Internet service in the municipalities, all located in the seven regions of the country in which USAID had ongoing programs. Does the availability of local telecenters encourage greater citizen involvement in politics? Here we look at how a randomized design could be planned to allow for rigorous evaluation of this question. In doing so, we see the usefulness of the intention-to-treat principle.

The Peruvian government required that telecenters be located in municipalities that lacked Internet service, so as to preclude competition with private providers. Around 200 municipalities in the seven regions in which USAID worked were eligible on this basis to participate in the program. Suppose that 100 of these 200 municipalities were selected at random and invited to participate in the program, while the other 100 municipalities served as controls.

At the end of the program, surveys of citizens living in both types of communities could record answers to questions about citizen involvement in politics. For example, the survey could record whether respondents had contacted a government official in the previous year. One possibility would then be to compare the political involvement of citizens in communities that have telecenters at the end of the program to those that do not. This can be misleading, however, because authorities in some communities assigned to the treatment group may refuse to participate in the telecenter program, and the same characteristics that influence whether local governments accept the telecenters may influence the degree to which citizens participate in politics.

Instead, we should compare the responses of citizens in the 100 communities that were invited to participate in the program—whether the communities ended up with telecenters or not—to the responses of citizens in the 100 control communities. This provides a valid basis for inferring the causal impact of the program, because confounding local characteristics should be independent of whether the community was invited to participate in that invitations are issued randomly. This intention-to-treat principle is important in many contexts in which USAID works, because programmers cannot tell in advance who will accept participation in the program.

Some hypothetical data may serve to illustrate the point. Suppose that at the end of the telecenter program, 10 people are sampled from each of the 200 municipalities at random, and that (for simplicity) each resident has an equal probability of selection into the sample. A survey is then administered to these people and citizens are asked about having contacted the government in the previous year.

If, hypothetically, 400 citizens in the intention-to-treat group had contacted the government, while only 250 citizens among the control group did so, we could estimate that random assignment to treatment raised individual responses by an estimated 60 percent.

$$\frac{400}{250} = 1.6$$

More complicated examples might call for more complicated estimators. Yet the intention-to-treat principle always provides a robust method for using experimental designs to evaluate whether interventions had a causal impact. In this hypothetical example, the experimental evaluation demonstrates that the telecenter program increased citizen participation in politics.

[Note: the intention-to-treat analysis may result in a dilution of the treatment effect since it ignores the fact that some citizens in the intention-to-treat group live in municipalities that refused treatment. An alternative in this context is to estimate the "effect of treatment on the treated".

Some units must be treated for political or other reasons. For political or other reasons, allocating treatment to one or several specific units may sometimes be nonnegotiable. For instance, political considerations sometimes come from host government institutions, and sometimes from U.S. foreign policy concerns. For programming evaluation purposes, the best solution is to implement the activity in these nonnegotiable units, then randomize other eligible units to treatment and control. The key is that for purposes of drawing conclusions about program impacts, outcomes should be compared across the randomly assigned units. The nonrandomly selected units that had to be treated for political reasons should be left out of the comparison.

One can always look as well at outcomes in the nonrandomly selected units. But comparing outcomes in such units to nontreated units will be less informative about the causal impact of the USAID intervention than comparing outcomes across the units that were randomly assigned to treatment and control. This happens because the nonnegotiable units may have differed from other units in the first place, in ways that matter for the outcome in question.

A slightly different issue is that host governments and other actors are sometimes compelled to "put out fires" during treatment. For example, while evaluating the impact of municipal-level interventions in mining towns on the likelihood of company-community conflict, the national government may be compelled to intervene when conflict breaks out in a community—whether that community is in the treatment group or among the controls. This might seem to pose important inferential issues. However, such "fires" may be as likely to occur in treated as in control units because treatment is randomly assigned. If treatment is randomized in large geographic clusters, this may be a more important issue, since exogenous factors like economic crises, political and social disturbances, and national disasters may have varying impacts in different regions. Yet, to the extent that there are reasonable numbers of treatment and control units, and to the extent like units are less geographically clustered with like units by design, this will tend not to pose an important issue for experimental inference.

All units are treated. In some decentralization programs, it may be the case that all eligible subnational units—municipalities or provinces, for example—must be treated, either because of political considerations or because the program continues a previous program in which an experimental design was not considered. This constraint may be more apparent than real, since it may be possible in some instances to create a true control group by randomizing some units out of treatment.

If such a control group cannot be established, experimental evaluation may still be possible. One useful strategy would be to randomize one set of units to one treatment and another set of units to a different treatment. Such an approach would not allow us to learn the average impact of treatment compared to no treatment, but it would allow us to estimate the marginal effectiveness of one treatment relative to a different treatment. For example, it might allow us to compare the impact of relatively top-down and bottom-up methods of local decision making, or to bundle different kinds of interventions and assess their joint impact.

Ethical implications. Ethical questions may sometimes arise in connection with experimental evaluations. For instance, is it ethical to deny treatment to control groups? One consideration is that we often do not know ahead of time how much benefit a USAID intervention will bring. Without experiment, we cannot know what will work and what will not. Well-designed, rigorous experimental protocols will allow knowledge to accumulate and thus enable decentralization programming to be as effective as possible in enhancing stability, promoting democracy, and fostering economic development.

In addition, there are virtually always untreated units in decentralization interventions—at least in working in the subnational arena and often in working with CSOs. For example, in the context of a decentralization program, it is often infeasible for USAID to work with all municipalities. The question becomes how the untreated units will be chosen. As discussed below, in many contexts it may be fairest and most ethically defensible to choose treated and untreated units randomly.

Envy of untreated units. How can implementers manage tensions arising from the envy of nontreated control units? First, the experimental design itself can help decrease the probability that authorities in control units become aware of treatments administered to neighboring, treated units. For example, neighboring municipalities or provinces can be clustered, and randomization to treatment or control can take place within the cluster.

Second, as discussed above, some experimental designs involve not an absence of treatment in control units but rather the implementation of a different treatment. One possibility is to administer one bundle of interventions to one set of municipalities and another bundle of interventions to a second set. Again, randomization of the bundles to municipalities could take place at the regional or subregional level. Though intermunicipal and interprovincial meetings might provide the opportunity for learning about differences across groups, these differences may not seem politically important, since all municipalities are receiving a treatment.

A third approach rests not on the particular experimental design but on the ethical implications of randomization. There are situations in which it may seem fairer to use a lottery to randomize units out of treatment. For example, in a follow-on decentralization program in a given country, fewer municipalities may be involved than the first phase. Is it not most politically palatable to tell municipalities no longer receiving assistance that the municipalities were chosen for the second phase by lottery?

Other donors flood the controls. One inferential issue relates to donor coordination (discussed in Section 5.1.6). It sometimes occurs that other donors may concentrate their programs in areas in which USAID does not work. While, as a programming matter, this can create valuable complementarities between the work of USAID and other donors, it can also complicate M&E tasks if other donors focus their work on the control units in a randomized evaluation of a USAID intervention.

One possible solution to this problem is to coordinate with other donors on a general geographic area or areas in which USAID will work—before USAID selects subnational units for treatment and control within these areas. Another solution would be for USAID to treat all subnational units in the area in which it works but randomize them to different sets of interventions. In other words, USAID would work with all municipalities in a predetermined set of regions, obviating the concern about other donors flooding the controls, but would randomly assign different treatments to different municipalities.

Absent either of these solutions, if donors focus on the control units, then comparing outcomes in the subnational units in which USAID works to the control units in which other donors work allows an estimate of the marginal effect of USAID's programming relative to the programming of other donors.

Contamination. This is an inferential, not a political, issue. In standard models of experimental inference, the response of one unit to treatment is not affected by the response of other units. Violations of this assumption can be problematic for inference. This is likely to arise as an issue in some experimental evaluations of decentralization programming. However, design modifications can often help. For example, if the intervention involves trainings or workshops with municipal mayors, randomizing at the provincial or regional level—say, subjecting all municipalities within a given province to either the treatment or control condition— might decrease the probability that mayors are aware of treatments administered to neighboring municipalities.

Gathering outcome data on controls. One challenge in the context of randomized trials—as well as other M&E plans in which treated and untreated units are compared—involves the gathering of outcome data on control units. In some contexts, gathering data on control units may be relatively straightforward, such as when citizens in both treated and untreated subnational units are surveyed or when experts provide evaluations of municipal capacity. In other contexts, it may be much more difficult, since USAID data on treated units are often gathered in the context of ongoing relationships between program implementers and local authorities. However, much of the difficulty of gathering data on controls probably relates more to output measures (not extensively discussed in this chapter) than to the outcome measures of interest here. Gathering output measures, such as the number and kind of meetings attended by local authorities, may indeed be difficult without the inducement provided by program participation, but these are not the indicators of greatest interest to M&E geared toward attributing program impact.

Heterogeneity of treatment effects. There often may be substantial heterogeneity of treatment effects across municipalities. For instance, in a decentralization program, interventions might have a big impact in some localities and negligible impact in others. Experiments help us estimate the average response to treatment across all units but may not help us assess the heterogeneity in treatment effects.

However, if we have strong reason to expect that the effects will be different in different subgroups, and we have a reason for wanting to assess these differences, then randomization among subgroups would help us estimate the impact in distinct kinds of municipalities. For example, one could randomize treatment within more densely populated urban districts and within more sparsely populated rural districts.

6.2.3 Nonexperimental Evaluations

While experiments can offer a feasible evaluation technique in some instances and are particularly useful for evaluating decentralization programming, they will not be available in all contexts. It is therefore important to consider nonexperimental designs as well.

For some decentralization interventions—in particular those where the intervention takes place in the subnational or civil society arenas and it is possible to compare a relatively large number of treated and untreated units—good nonexperimental evaluations will share some of the same basic features as experimental designs. First, it is crucial to find or construct good outcome indicators, rather than use indicators that only monitor the performance of local partners or track the process of program implementation (see Section 6.1). Second, in programs where a relatively large number of units are treated, it is essential to gather data on control units. For purposes of estimating when the number of units is large enough for some comparisons, explicit ex ante calculations of statistical power may be useful.

Consider a nonexperimental evaluation of a program, the strategic objective of which is to increase the ability of CSOs to partner with subnational governments and administrations. USAID and its partners may work with local CSOs in a selected number of municipalities or regions but not others. Following the proposed outcome indicators in Table 6.1, program evaluators might gather data on the number of public-private partnerships for infrastructure projects and service delivery, or on the extent of private investment in employment-generating projects. These data should ideally be gathered—at a minimum—before the intervention and at the end of the program.

Crucially, the data should also be gathered on both treated and untreated units. Consider the alternative, which is to simply gather baseline data on treated units before and after the program. In this case, the observation that the number of public-private partnerships or the extent of private investment has increased or decreased over the life of the program is very difficult to attribute to the impact of the program, since many other factors may have also varied over time, influencing the change in outcomes. Suppose, for instance, that the economy grew rapidly in treated units. Then changes in the extent of private investment during the program could well be due to these changing economic conditions, rather than to the effect of USAID's work with local CSOs. Many other examples make clear that it is far more powerful to compare treated to untreated units. The best comparisons might use pre- and post-intervention data to compare treated and untreated units over time—for example, by comparing the change in the extent of private investment in treated units to the change in the extent of private investment in untreated units.

The question then becomes how best to select untreated units for comparison with treated units. In general, the best approach is to pick untreated units that look as much as possible like the treated units. Of course, this is what randomization accomplishes: on average, randomization balances third factors that might account for different outcomes across the treatment and control groups, thus allowing us more confidently to attribute differences across the groups to the impact of the intervention.

When actual randomization is not available, perhaps because the program evaluation was not well designed before the program or because the nature of the program changed substantially due to extenuating circumstances, there may still be other good alternatives. Probably the best approach in the context of many decentralization programs is to pick a subset of treated and untreated units to compare with the nonrandomly selected treated units, where the subset would be matched as closely as possible with the treated units in terms of variables that might affect the outcomes of interest.

For example, it may be possible in some decentralization programs to compare treated subnational units on one side of a jurisdictional border to similar untreated units on the other side of the border. Suppose decentralization interventions are undertaken at the municipal level in several (nonrandomly selected) provinces or regions of a given country. Even though the regions themselves were nonrandomly selected, there may be a credible case that municipalities on one side of a provincial border—say, municipalities that are inside the treated provinces or regions and thus are subject to the interventions—are similar with respect to potential confounding factors to municipalities that are on the other side of the border and thus not subject to the intervention.

To continue the example above, it might be the case that economic conditions changed over the life of the program in the subset of treated units on one side of the border, but they may also be expected to have changed in nearby municipalities just on the other side of the border who were not subject to the intervention of interest. Differences in the extent of private investment across these treated and untreated units can then be more confidently attributed to the effects of USAID's work with local CSOs. Even better, one might compare the difference over time or change in outcomes in the treated and untreated municipalities. Even if there are preexisting differences in the level of private investment in the treated and untreated units, comparing the change in investment in the two groups of units will cause this initial difference to wash out.

Finally, it is useful to say something about interventions that may not admit evaluations of the kind discussed thus far. While many of the interventions undertaken in connection with decentralization programming are amenable to comparisons across treated and untreated units, some are not. For instance, interventions in the national arena will tend to be much less amenable to such comparisons than interventions in the subnational or civil society arenas, often simply because there is only one unit—the national legislature, say—with which USAID and its partners will work.

In such cases, other kinds of evidence might be useful. There is a long social scientific tradition, for example, that emphasizes the use of counterfactual reasoning, as well as modes of causal inference in which key nuggets of information—causal "smoking guns," sometimes called causal process observations—can play a key role.

Suppose, for example, the question is whether advocacy by a USAID-supported CSO played an important or crucial role in the passage of a new national decentralization law. The key counterfactual question may be: would the law have been passed if the CSO did not exist or had stayed out of negotiations over the bill? Some pieces of information might be key to bolstering inferences about this question. For instance, political parties may have met at the offices of the CSO to negotiate approval of the law, or the CSO might itself have played an important role in writing sections of the draft legislation. Various sources, including interviews with involved actors, could provide such information. Of course, this approach may not develop beyond the anecdotal, and it seems difficult as a general matter to make recommendations as to how causal process observations should be found. In evaluating the effects of such interventions in the national arena, where the more robust methods on which we have focused are not feasible, inferences about the impact of programming are likely to be more tentative.

6.3 PRACTICAL ISSUES FOR SUCCESSFUL MONITORING AND EVALUTATION

This chapter has focused on evaluating the impact of USAID interventions in support of decentralization initiatives. The emphasis has been on the methodological issues involved in measuring outcomes and inferring program impacts. Yet many obstacles to successful monitoring and evaluation activities may be practical rather than methodological in nature. This concluding section focuses on three potential obstacles—incentives to do M&E for impact evaluation, skills and capabilities, and the costs of rigorous experimental and nonexperimental evaluations—and suggests strategies for overcoming or mitigating these obstacles.

6.3.1 Incentives to do M&E

Much M&E activity focuses on the process of program implementation. This sort of M&E can be crucial for field officers and implementing partners who are tracking progress of decentralization programming, and it can serve as an important program management tool. Given the many other demands on the time and resources of field officers and implementing partners, sometimes incentives to conduct this kind of M&E can be weak as well. But the immediate benefits in terms of program management may often appear to justify the costs of this kind of M&E activity.

With M&E geared toward impact evaluation, however, program effects may take some time to detect. There are also a number of different tasks that need to be implemented, including the randomized selection of units and the need to collect data on treated and control units. As a result, one might wonder whether the benefits outweigh the costs. The answer, generally, is an emphatic yes. Only through rigorous program evaluations is there any hope of getting solid evidence of program impact and of learning what programming initiatives work best. Over time, this kind of evidence can accumulate and allow programming to become more effective.

Yet it is important to be aware that there will be instances in which more rigorous methodologies reveal little or no program impact. Though disappointing, these findings are also useful for the aggregate goal of improving programming effectiveness. Without assessing this evidence, there will be little basis for assessing what has worked and what has not. It is therefore important that implementing partners and field officers not be held accountable for the results of impact evaluations. Rather, the emphasis should be on the gradual accumulation of knowledge through rigorous evaluations, and partners and field staff should be rewarded for the rigor of the evaluation, not its results.

To enhance incentives to carry out the rigorous impact assessments described in this chapter, it may be crucial to separate evaluation of program impact from the standard M&E tools that help to assess program implementation and that can serve as useful program management tools. For example, in performance monitoring plans, implementing partners may be held accountable for meeting targets—say, holding a certain number of meetings, workshops, and training sessions—for program implementation.

6.3.2 Skills and Capabilities

This chapter has introduced some basic ideas about rigorous experimental and nonexperimental designs that are useful for evaluating program impact. Some of the ideas are straightforward. But actually implementing such evaluations can demand expertise in research methodologies.

When designing and implementing program evaluations, it will be valuable to bring to bear the resources, skills, and capabilities present within USAID and partner organizations. USAID's Office of Democracy and Governance sponsors Democracy Fellows with expertise in evaluation methodology, and other experts may be available for trainings and consultations through USAID/DCHA/DG.

In addition, partner organizations, including local partners on decentralization projects, increasingly offer expertise in randomized evaluations. Building in requests for experimental evaluations and other rigorous designs at the Request for Proposal stage of a decentralization program can be a useful way to identify partners with the necessary skills and capabilities to implement these evaluation designs.

6.3.3 Costs of Rigorous Evaluations

The issue of cost raises several considerations for program evaluation. Relative to a standard but weak nonexperimental design, which is to gather baseline and post-intervention data on treated units, the major additional cost of implementing an experimental evaluation involves the need to gather outcome data on control units. For instance, surveys may need to be conducted in both treated and control municipalities, so that citizen evaluations of governmental responsiveness in treated and untreated units can be compared. In the case of surveys, adding additional respondents from control municipalities is often not very expensive and would typically represent a fraction of the overall cost of M&E for a given project. In the case of other kinds of outcome indicators, gathering data on controls could conceivably be more difficult or costly (see the discussion of gathering outcome data on controls in Section 6.2.2).

Often, without appropriate controls (whether the controls are chosen through random assignment or other means), little can be said about program impact. Gathering data on controls is thus essential for making meaningful assessments of program impact and should be carefully considered as an option for every M&E plan. It should also be noted that once data are going to be gathered on control units, the material cost of choosing the treated and control units randomly is usually next to nothing. Additional costs for M&E stem not from randomization but from the need to gather outcome data on controls.

Experimental evaluations may also present the opportunity for substantial cost savings. In principle, random assignment of units to treatment and control can do away with the need to gather baseline data on either treated units or the controls. With enough units, baseline characteristics will be approximately the same in both treated and untreated units, due to the randomization. (In practice, it can sometimes be useful to gather baseline data, both as a randomization check and as a useful source of information for subsequent subsample analysis.) There may be other opportunities for cost-saving associated with reducing emphasis on M&E for assessing the progress of program implementation—which often requires extensive record-keeping—and moving toward greater emphasis on M&E for assessing program impact. Other cost-effective measures—such as sampling subnational units for gathering outcome data, rather than gathering outcome data at several points in time in all units—can sometimes be deployed.

It is important to note that using experimental approaches often involves careful thinking about the selection of units and other design issues early in the program, as this is the point at which design choices can be made that will allow for rigorous evaluation of program impacts. With experimental approaches, some of the M&E costs may therefore come earlier in the process relative to other standard forms of M&E. To ensure that the M&E methods selected are both rigorous and cost-effective, it may be useful to consult persons with methodological expertise inside USAID and partner organizations when designing an M&E plan.

6.4 CONCLUSIONS REGARDING PROGRAM EVALUATION

This handbook discusses strategies to support decentralization activities that can help to enhance stability, promote democracy, and foster economic development. Yet some strategies may work better than others, and strategies may work differently at different times and in different places. The best way to evaluate whether and when such programming works to promote its objectives, is by designing and implementing rigorous monitoring and evaluation plans. This chapter has described some of the newer thinking about how best to conduct evaluations of program impact.

The chapter makes several key points. The first is that successful program evaluation involves, as a necessary first step, the definition of strategic objectives and outcome indicators. Strategic objectives may involve an aspect of the broad goals of enhancing stability, promoting democracy, or fostering economic development. For example, "reducing conflict or contentious actions" might be one strategic objective under the heading of promoting stability. Then outcome indicators, such as the number of marches or protests (perhaps per capita by jurisdiction) must be defined. This chapter provided some illustrations of strategic objectives and outcome indicators, but these are likely to be program and context specific. In practice, other measurement issues will arise as well, including units of measure, data sources, frequency of data collection, and the party responsible for data collection and reporting.

Once these objectives and outcome indicators are defined, the plan for evaluating program impact must be designed. The second main point of this chapter is that, where feasible, randomized evaluations provide the best possible strategy. Randomized evaluations offer the most reliable basis for assessing program impact, and decentralization programming is quite amenable to randomized approaches. Particularly with interventions in the subnational or civil society arenas, eligible subnational units or organizations should be identified and listed, and the size of the desired treatment group should be determined. Treatment and control units should then be chosen at random from the list of eligible units. Potential obstacles to random assignment—such as the idea that political will determines where programming interventions can take place or the concern that some units must be treated for political reasons—are important to consider, but these can often be countered by appropriate design modifications.

Finally, planning M&E activities—drawing on expertise provided by USAID offices and implementing partners—is key to enabling the accurate assessment of decentralization programming. Fundamental here are the selection—before program implementation—of treatment and control units, and a plan to gather data on both groups. This is crucial for effective and rigorous evaluation. Only by conducting rigorous evaluations can knowledge about the impact of decentralization programming gradually accumulate and serve its ultimate purpose: to improve the quality of programming in support of decentralization initiatives, thereby helping to enhance stability, promote democracy, and foster economic development.

Questions for Review – Chapter 6

You may test and reinforce your understanding of selected concepts presented in this chapter by responding to the following questions before reading the concluding chapter. Answers to questions will be found on the pages indictated with each question.

1. ***What are the advantages of measuring program outcomes rather than program outputs? (p. 76)***
2. ***Describe why randomized trials are effective ways to assess program impact. (pp. 80 - 82)***
3. ***In a country with which you are familiar, how would you deal with randomized trials given the concerns of:***

a) ***Political will? (pp. 82 – 83)***

b) ***Ethical implications? (p. 85)***

c) ***Jealousies of untreated units? (p. 85)***

4. ***Using Table 6.1 as guide, develop objectives for each of the three main goals of stability, democracy, and development for the subnational arena in a country with which you are familiar. You also may use objectives that you have already developed. (pp. 77 - 79) Using the objectives you developed, create at least two outcome indicators for each objective (pp. 77 - 79).***

5. ***For each outcome indicator you developed, describe how each could be evaluated using randomized trials. (pp. 80 - 82)***