National Academy of Sciences
USAID Program Evaluation

Peru Field Visit

July 15, 2007

Thad Dunning -- Field Memos

Index of memos

1. Overview: Improving USAID program evaluation

How can the impact of USAID democracy and governance programs be better evaluated?

These memos provide several suggestions, drawing on experiences during a recent field visit to Peru. In the memos, I discuss how experimental (randomized) designs could have been introduced in specific programs already completed, and I describe the potential for experimental evaluation of several programs currently being designed. I also consider non-experimental evaluation of interventions that may lend themselves less readily to randomization as well as the possibilities for retrospective evaluations of past programs' impact.

The main conclusions are as follows. First, the introduction of experimental designs into USAID program evaluation, where appropriate, would represent an improvement over current practice. I discuss below aspects of a number of program monitoring plans in Peru that could have been or could be improved by experimental evaluations, including principally the decentralization program but also interventions relating to the rule-of-law, political parties, extractive industries, and rural connectivity.

Second, experiments are both feasible and cost-effective in many contexts. Our field experience in Peru suggests that many objections to the feasibility of experimental evaluation can often be countered by design modifications or other approaches. I discuss some general ideas in a section of the second memo on "standard objections to experimental designs." One key concept is the intention-to-treat principle, which may help to counter the common notion that the dependence of program implementation on "political will" makes experimental design infeasible.

Third, however, more elementary improvements to USAID evaluation practices may be likely to bring bigger marginal gains for attributing program impact. For example, the program monitoring plans we examined in Peru required the reporting of indicators that are largely related to contractor performance and very proximate "outputs" of the intervention, like the number of meetings held or trainings conducted. These indicators, however, are largely useless for purposes of evaluating the impact of programs on broader outcomes of interest. While USAID documents and our discussions in the field suggest that the agency staff are well aware of the distinction between monitoring contractor performance and evaluating program impact, in practice the scales seem clearly tipped towards the former activity.

In addition, there is a near-absence in the programs we examined of comparisons across units subject to USAID interventions (the treated units) and untreated units, that is, controls. Even when interventions would readily lend themselves to the comparison of treated and untreated units, as in the decentralization program, there appears to be very little effort to gather data on (appropriately-chosen) control units, or even awareness of the utility of doing so. In many cases, as the discussion below suggests, gathering such

data need not represent a great additional cost; with appropriate design modifications, cost savings may even be available.

Finally, it may be possible to draw on accumulated data to conduct informative retrospective evaluations of program impact. For such efforts to be useful or convincing in assessing program impact, however, it appears that in many cases it would be necessary to gather additional data, for example on appropriately selected control units. It may be worthwhile for USAID to commission studies that would gather the appropriate data and conduct analyses along the lines suggested below.

In sum, from the perspective of impact evaluation, improving outcome measurement and gathering data on untreated units represent the most important recommended changes to current USAID practice (at least as current practice is reflected in Peru). Absent randomization of treatment, of course, comparisons across treated and untreated units may be seriously misleading for purposes of drawing causal inferences about the impact of treatment. Yet improving outcome measurement and gathering data on controls, whenever the nature of the program allows, seem necessary first steps, and of course experimental approaches cannot work unless both of these initial improvements are first put into place.

These conclusions will not apply to all USAID programs or interventions; some interventions will not be amenable either to experimental designs or the kinds of non-experimental approaches I discuss. Yet we found in Peru that such design modifications could help improve the evaluation of the impact of many programs. Improving program evaluation in this way could assist with the overall goal of helping USAID know with greater confidence whether programs have causal effects and what those effects may be.

In the rest of this overview memo, I reflect further on what light our Peru field visit can shed on important themes considered by the National Academy of Sciences committee.

1.1 Operational Definition of Democracy

USAID approached the National Academy of Sciences to seek advice on assessing the impact of AID democracy and governance programs. One outcome variable of interest involves what committee members have called "big D" democracy: the nature of the political regime at the national level. The committee has proposed consideration of thirteen "dimensions" of democracy, three of which -- transparency, civil society, and sub-national government -- were chosen for closer scrutiny during field visits.

In my own view, such variables are probably best conceptualized not as dimensions of democracy but as variables that may or may not have causal relationships to the political regime (as causes, effects, or both). One classic defense of procedural-minimum approaches to democracy, of course, is that restricting the components of the core concept (to, say, participation and contestation, as in Dahl's classic formulation) allows analysts to assess hypotheses about the *relationship* of variables like civil society or sub-

national government to the political regime. Building such variables into the definition of democracy does not appear to allow us to test such hypotheses.

This being as it may, assessing hypotheses about the political impact of civil society organizations or decentralization seems very much in the spirit of the committee's draft document (CEUDAP 2007a). For example, the presence of "dynamic, independent, and politically active" groups may or may not be a good thing for democracy.[1] Whether decentralization of political power increases or decreases ("big D") democracy is also an empirical question.[2]

These hypotheses are also, of course, of great interest to USAID, because the impact of USAID assistance on regime outcomes must ultimately run through a causal chain, such as: (1) USAID assistance fosters a more "robust" civil society; and (2) a more robust civil society in turn promotes democracy.

Unfortunately, it is very difficult to assess hypotheses such as (2) using within-country evidence, at least when employing the kinds of research designs I mainly consider in these memos. For one, there is only one aggregate regime outcome at any point in time, and this outcome is obviously over-determined with respect to the variables of interest. For another, there is no control unit appropriate for investigating empirically the key counterfactual question: what would have happened to democracy in the absence of a robust civil society?

The kinds of hypotheses we considered in Peru therefore relate to causal claims like (1): that is, they relate to the impact of USAID assistance on outcomes that are distant from the intervention but that nonetheless are viewed as intervening steps towards a larger goal. In the language of USAID strategic assessments, these are sometimes called "intermediate results." Outcomes of interest may also include broader "strategic objectives," such as "increasing the responsiveness of local governments to citizen demands," that often themselves constitute steps towards a larger goal, such as "building sustainable democracy" (USAID 1998).

Fortunately, it appears that there is huge room for improvement in how USAID evaluates the impact of its programs at these intermediate levels. Our experience in the field suggests that the importance of improving the evaluation of mid-range hypotheses should not be lost, even as one also keeps in mind the larger questions – say, the impact of USAID programs on "big D" democracy. Answering meso-level questions may be

[1] Consider the "dynamic, independent, and politically active" groups currently pressing for greater regional autonomy in the relatively wealthy Bolivian lowlands. Some analysts characterize the efforts of such groups as a backlash to the rising political power of indigenous majorities in the highlands (Eaton 2007) -- and thus arguably an attempt to subvert "rule by the people," which comprises the core component of the concept of democracy proposed in CEUDAP (2007a).

[2] The draft committee document asks: "How decentralized is political power and how democratic is politics at subnational levels?" (CEUDAP 2007a). However, positing that decentralization is a dimension of democracy would seem to limit the ability to ask causal questions about the impact of decentralization on the political regime. In addition, asking whether politics are democratic at sub-national levels seems to beg the question of the definition of democracy.

necessary if not sufficient for assessing the impact of USAID on the aggregate questions of interest. Yet in addition, there appears to be such a large leap between the proximate questions USAID can currently answer with confidence and the aggregate questions of interest that the importance of rigorous testing of meso-level hypotheses should not be discounted.

In sum, the focus on meso-level hypotheses does not discount the importance of investigating USAID's impact on broader outcomes like democracy. In the conclusion, I offer a few further thoughts on the issue of aggregating insights gained from program evaluation at this intermediate level.

2. Experimental designs

An *experimental design*, as the term will be used in this memo, is one in which (1) some subjects or units are invited to receive an intervention or treatment while others serve as controls; and (2) invitation to receive the treatment occurs at random, for example, through a lottery.

Both characteristics play an important role in experiments. First, average outcomes can be compared across the treatment group and the controls. Second, any differences bigger than what might reasonably have occurred by chance can be reliably attributed to the causal effect of the intervention. One reason is that confounding factors should not be responsible for any differences across the groups: random assignment ensures that other factors influencing outcomes should be balanced, on average, across the treatment group and the controls.

USAID programs typically involve an intervention or set of interventions, which may be implemented by local contractors. For example, in a decentralization program (to be discussed below), contractors may train local mayors and non-governmental organizations in "participatory budgeting," in which citizens and elected authorities work together to formulate local fiscal plans; they may also conduct workshops on the local implications of a national decentralization law, provide technical assistance to civil society groups, and implement other interventions. The goal of program evaluation should be to assess whether these interventions have a causal impact on outcomes of interest – in the case of the decentralization program, the responsiveness of local governments to citizens' demands.[3]

What is an experimental design in this context? The core ideas are straightforward, and they parallel general features of experimental designs. In the case of decentralization, some municipalities may be invited to participate in trainings, workshops, and other interventions (the treatment group), while others are not (the controls). The key is that invitations to participate are issued to municipalities at random. One can then attribute differences in government responsiveness to the effect of treatment with greater confidence than in non-experimental designs.

A third standard feature of experimental designs, not mentioned above, is that the experimental intervention is manipulated. In a medical trial, for example, researchers may dispense pills to the treatment group (and perhaps placebos to the controls).

How should we think about the manipulability of interventions, in the context of USAID programs? In some USAID programs, the nature of the interventions may be quite vague. Yet there is virtually always some manipulation. In a given program, USAID allocates money towards some strategic objective and hires US and local contractors and

[3] Challenges involved in measuring the outcome are discussed below.

grantees to undertake a set of activities related to that objective. Those activities constitute the manipulation or manipulations of interest.

This does not imply that all units assigned to the treatment group are actually subject to the manipulation, however. In a decentralization program, for example, some mayors will not sign participation agreements or otherwise cooperate with contractors to ensure that trainings and workshops occur.

These non-cooperating units can be likened to non-compliers in a medical trial – those patients assigned to treatment who do not take the pill. Note that the presence of non-compliers does *not* imply that the treatment was not manipulated or is not potentially manipulable. It does, however, have important implications for analyzing the results of experiments.

2.1 The "intention-to-treat" principle

A standard approach to analyzing experiments in which there is non-compliance involves the intention-to-treat principle. This principle is important to keep in mind in the context of USAID program evaluation, for two reasons.

First, intention-to-treat analysis provides a robust basis for causal inference in the presence of non-compliance, which is endemic in USAID programs. Second, this principle may help counter many standard objections to experimental designs in the context of USAID programs -- for example, that experimental designs are impossible to implement because of the role of "political will."

With respect to the first point, consider an analogy to a medical trial designed to assess the impact of mammography on death rates from breast cancer. A problem is that not all women invited to come in for screening actually show up. Comparing death rates from breast cancer among women who showed up for screening and those who did not can be misleading: women self-select for screening, and confounding factors (like health consciousness) may be associated with both seeking screening and lower death rates from breast cancer.

The intention-to-treat principle says that we should compare the death rates of women who were randomly *invited* for screening to those who were not invited; since the invitation is randomly assigned, it will be independent of confounding factors.[4] Intention-to-treat analysis therefore provides a robust way of assessing the causal impact of screening (for details, see Freedman, Petitti, and Robins 2004).

In the context of a USAID decentralization program, we should compare municipalities randomly *invited* to participate in trainings and workshops to those not so invited. We should *not* compare municipalities that actually participated to those that did not:

[4] Note that some women invited for screening (the treatment group) will not come in for mammographies; but a roughly equivalent proportion of women who were not invited (the controls) also would not have come it, had they in fact been invited.

municipal authorities who agree to participate may be different than those who do not, in ways that matter for outcomes such as government responsiveness to citizens' demands.

However, confounding municipal characteristics should be independent of whether the municipality was *invited* to participate, which makes intention-to-treat analysis a robust way of assessing whether there is a causal effect. I discuss the details further below. (It may also be useful to consider the "effect of treatment on the treated," discussed in the Appendix).

The intention-to-treat principle may be useful to keep in mind for a second reason as well. One of the seeming obstacles to experimental evaluation of USAID programs, frequently cited during a recent field visit to Peru, is that contractors can only work with local actors when there is the "political will" to do so.

In an intention-to-treat analysis, however, only some of the municipalities invited to participate will possess the appropriate degree of "political will," while others will not.[5] Municipalities in the treatment group that lack political will are like non-compliers in a medical trial. Yet as discussed above, non-compliance does not pose an obstacle to experimental analysis. For practical purposes, the key is to have a set of units that are randomly assigned to be invited for treatment, and another set of units that will not be invited. USAID or its contractor can then seek to work *only* with units in the first group.

The intention-to-treat principle is therefore quite useful for countering many standard objections to experimental designs that we encountered in the field. (See the section below on "standard objections to experimental designs"). As I discuss further below, the intention-to-treat principle may be a useful concept for experimental evaluation of USAID programs related to rural connectivity, extractive industries, or the strengthening of political parties.

2.2 Case studies

Not all USAID programs or interventions will be amenable to experimental design. At a minimum, sufficient numbers of units must be available to provide a treatment group and and controls. (How many units must be available for a useful experimental design depends on a variety of factors, discussed below). I discuss non-experimental program evaluation in other memos.

Yet the sense of the National Academy of Sciences committee, and my sense after field visits to Peru, is that USAID program evaluation could be improved in many instances by introducing experimental designs. In this section, I discuss a series of interventions in USAID/Peru programs that were *not* evaluated experimentally but that could have been. These case studies may help to clarify the points introduced above and will also introduce new issues.

[5] A roughly equivalent proportion of control units will possess the appropriate degree of political will. This is because the randomization ensures that assignment to treatment is independent of political will.

It may be useful to keep in mind that USAID Democracy and Governance programs usually contain a broad mix of interventions. In the programs I examined in Peru, it was the case that some interventions in the program were amenable to the introduction of experimental designs, while others are not.

Finally, it may be worth keeping in mind a general point made in the overview memo. Experimental designs would often be useful for USAID's program evaluations. Yet the issues raised by non-random assignment for current program evaluation are probably dwarfed in importance by issues related to the measurement of outcomes and the near-absence of control units (randomly selected or not). These issues certainly must be resolved before any experimental approach is viable.

2.1.1 Decentralization

USAID/Peru launched a program in 2002 to support national decentralization policies initiated by the Peruvian government. Over a five-year period, the Pro-Decentralization (PRODES) program was intended to

- support the implementation of mechanisms for citizen participation with sub-national governments (such as "participatory budgeting");
- strengthen the management skills of sub-national governments in selected regions of Peru; and
- increase the capacity of non-governmental organizations in these same regions to interact with their local government (USAID/Peru 2002).

With the exception of some activities relating to national-level policies, all interventions under the program took place in seven selected sub-national regions (also called departments): Ayacucho, Cusco, Huanuco, Junin, Pasco, San Martin and Ucayali.[6]

These seven regions contain 61 provinces, which in turn contain 536 districts.[7] Workshops on participatory budgeting, training of civil-society organizations, and other interventions took place at the regional, provincial, and district level.[8]

The ultimate goal of the program was to promote "increased responsiveness of sub-national elected governments to citizens at the local level in selected regions" (USAID/Peru 2002). This outcome is potentially measurable on different units of

[6] As discussed elsewhere, the regions were non-randomly selected for programs because they share high poverty rates, significant indigenous populations, narcotics-related activities, and because a number of the departments were strongholds for the Shining Path movement in the 1980's.

[7] Peru has 24 departments plus one "constitutional province;" the 24 departments in turn comprise 194 provinces and 1,832 districts. Provinces and districts are often both called "municipalities" in Peru and both have mayors. Sometimes two or more districts combine to form a city, however.

[8] Relevant sub-national authorities include members of regional councils, provincial mayors, and mayors of districts.

observation. For example, government capacity and responsiveness could be measured at the district or provincial level (through expert appraisals or other means), while citizens' perceptions of government responsiveness may be measured at the individual level (through surveys).

In the rest of this case study, I first describe the actual program evaluation plan that was implemented and offer estimates of the cost of monitoring and evaluation under the actual plan. I then comment on "best-possible designs" that would have offered improvements over the actual design, from the perspective of evaluating program impact, and estimate the costs that such a design might have involved. I suggest that experimental designs could be used to study the impact of the decentralization program, and that the cost of appropriately-designed experimental evaluations could in fact be far beneath the actual costs spent on monitoring and evaluation.

Because the USAID/Peru decentralization program is now entering a second five-year period, which offers opportunities for revamping program design, I offer some suggestions on improved design moving forward as well.

Actual program monitoring plan

The PRODES decentralization program represented an ambitious effort. By all accounts, it was a well-executed program; the performance of the local contractor received high marks from Mission staff at USAID/Peru.

The questions of interest here do not relate to the performance of the contractor, however. Instead, the question is how we can know whether such a program had an impact on larger outcomes, such as the responsiveness of local governments to citizens' demands.

For purposes of impact evaluation, the design of the decentralization program suffered from a number of serious deficiencies. With foresight, some of these deficiencies might have been fairly easily corrected, and not at much additional cost; in fact, cost savings were likely available. The design of such a program could certainly be improved to allow for a better assessment of impact.

The main deficiencies I note here parallel the general points raised in the overview memo above: absence of outcome indicators, absence of control units, and absence of treatment randomization. The first two points do not relate solely to experimental design; yet they are certainly necessary for experimental approaches to be useful, so I discuss them first before turning to the issue of randomization.

Outcome measurement. As mentioned, the decentralization program sought to foster citizen participation, transparency, and accountability at the local level, with the ultimate objective of promoting "increased responsiveness of sub-national elected governments to citizens." Though some of these outcomes are potentially, albeit imperfectly, measurable, indicators gathered at the local level related almost exclusively to what might be called *outputs* rather than outcomes.

For example, gathered indicators included:

- the percentage of municipalities that signed "participation agreements" with local contractors;

- the percentage of participating municipalities from which at least two individuals (local authorities or representatives of civil society groups) attended a training course in participatory planning and budgeting;

- the percentage of targeted provincial governments in which at least two civil society organizations exercised regular oversight of municipal government operations, as measured by participation in at least two public fora during the year; and

- the percentage of participating local governments that establish technical teams to assist with decentralization efforts (USAID 2007).

Such indicators seem designed to monitor contractor performance and perhaps measure very proximate outcomes, such as formal participation in the decentralization process. However, they do little to help discern the impact of interventions on outcomes such as the responsiveness of sub-national elected governments to citizens.[9] [10]

Several indicators gathered through surveys did tap citizens' perceptions of the responsiveness of sub-national elected governments in targeted municipalities. Surveys taken in 2003, 2005, and 2006 asked respondents

- Are the services provided by the (district, provincial or regional) government very good, good, average, bad, or very bad?

Another question, administered only in the 2003 and 2005 surveys, asked:[11]

- Do you think that the (district, provincial, or regional) government is responsive to what the people want almost always, on the majority of occasions, from time to time, almost never, or never? (2006, 2007 PMPs)

In principle, such survey questions may provide useful proxy measures of the outcomes of interest. In practice, there were a number of issues that limited the usefulness of these measures. First, only the first question was asked in a comparable manner across the

[9] Outcomes measures were also only gathered on participating municipalities, as discussed below.

[10] The USAID/Peru team and local contractors were clearly aware of the distinction between measures of contractor performance and measures useful for assessing impact; this distinction is made in some of the relevant program monitoring plans (e.g., PRODES PMP 2006). However, most of the impact measures appear to be fairly proximate measures related to the process of supporting decentralization.

[11] The 2003 and 2005 questions were administered as a part the Democratic Indicators Monitoring Survey (DIMS), while in 2006 data come from the Latin American Public Opinion Project (LAPOP).

three surveys, allowing for a very limited time series on the outcome of interest. Second and perhaps more important, as discussed further below, was the failure to gather measures on control units in all but the 2006 survey.

Finally, a "baseline" assessment of municipal capacity was prepared at the start of the program by a local institution. All district and provincial municipalities in the seven selected regions were coded along several dimensions, including extent of socio-economic needs and management capacities of district and provincial governments (GRADE 2003).

Poverty rates and related indicators played a preponderant role in the local institution's calculations, which may have limited the usefulness of the index for assessing changes in sub-national government capacity or responsiveness. In theory, however, repeated assessments of this kind could have provided useful data on municipal capacity, which is an outcome of interest under the decentralization program. As far as we can tell, the assessment was not repeated.

Granted, some of the stated objectives of the decentralization program – including citizen participation and especially government transparency and accountability – are notoriously difficult to measure, and a priority should perhaps be to help USAID think about how to create indicators for difficult-to-measure outcomes. Yet improvements in the measurements of outcomes are certainly available, and this should be a first-order concern for improvement of program evaluation.

Absence of controls. USAID/Peru's contractor was tasked with implementing the decentralization program in all 536 districts of the seven selected regions. Once the roll-out of interventions in all municipalities had been completed, no untreated municipalities remained available within the selected regions.

The absence of appropriate control units (untreated municipalities) is perhaps the biggest problem for effective evaluation of the decentralization program. Since roll-out was completed by the second year of the program, there was little opportunity to compare outcomes in treated and untreated units within the seven regions.[12]

In principle, comparisons could be made across treated municipalities in the seven selected regions and untreated municipalities outside of these regions. Since the seven regions were non-randomly selected on the basis of characteristics that almost surely covary with municipal capacity and sub-national government responsiveness (e.g., high poverty rates, narcotics-related activities, and past presence of the Shining Path),

[12] The 2006 PRODES performance monitoring plan notes that the survey taken in 2003 allowed comparison of the "approximately 145 municipalities that were already incorporated into the program as well as the sub-sample - approximately 392 additional municipalities - that were yet to be phased in." This is the closest approximation to control units in the actual design, but the remaining municipalities were brought into the program in the following year.

inferences drawn from such comparisons would be problematic yet potentially could be informative.[13]

In practice, however, the data do not exist for such comparisons because virtually no data were gathered on control units. The exception is the 2006 commissioned survey taken as a part of the Latin American Public Opinion Project (LAPOP), which administered a questionnaire to a nationwide probability sample of adults including an over-sample of residents in the seven regions in which USAID works (LAPOP 2007).[14] This survey includes several questions that would be useful measures of the outcome variables (though only one question is comparable to questions asked in the earlier non-LAPOP surveys taken in treated municipalities in 2003 and 2005).[15]

Outside of the LAPOP survey, no data were gathered on untreated municipalities. The universe of the 2003 and 2005 surveys was limited to residents of the seven regions (and thus only to residents of treated municipalities). Evaluations of municipal capacity (e.g., the GRADE study mentioned above) were conducted only on districts and provinces in the seven selected regions.

The inferential issues here are obvious. As just one example, many municipalities in the seven regions had been ravaged by the conflict with the Shining Path during the 1980's and 1990's. Investment and population return has picked up in some areas during the last decade and especially the past five years; much of this upturn must be due to the end of the war and other factors.[16] Improvements in measured municipal capacity or in citizens' perceptions of local government responsiveness during the life of the program may therefore not be readily attributable to USAID support for decentralization.

In sum, as discussed further below, a plan for gathering data on control units should have been created ex-ante. Ideally, one would compare treated and untreated municipalities *inside* the seven regions. Absent the existence of untreated municipalities inside the regions, data could be gathered on appropriately-selected municipalities outside the region.[17] Surveys should have included residents of untreated municipalities, and evaluations of municipal capacity (such as the GRADE study) should have included pre- and post- measures on municipalities with which USAID/Peru's contractor was *not* assigned to work.

---

[13] They also almost certainly would not cast a favorable light on the impact of the decentralization program, since municipalities in the selected regions are likely to be less capable and less responsive to begin with. Pre- and post- comparisons across treated units in the seven regions and a subset of untreated units outside of the regions might be illuminating; see the memo on retrospective evaluations.

[14] In addition to 1500 respondents in the nationwide sample, an over-sample of 2100 (300 per region) was taken from the seven regions (Patricia Zárate, Instituto de Estudios Peruanos, personal communication). *Inter alia*, this survey asked respondents their opinions of the quality of local government services, as noted above.

[15] The LAPOP instruments include questions that are comparable across twenty surveyed countries; see LAPOP (2006). For useful information, we are grateful to Patricia Zárate, Instituto de Estudios Peruanos.

[16] Interviews, Ayacucho, June 27, 2007.

[17] However, as discussed below, data on controls may also not help with the inferential issues mentioned in the previous paragraph, absent random assignment.

It might also be noted in passing that an over-sample from the seven focus regions, taken as part of the LAPOP/PRODES survey, may not help much for present purposes. An over-sample would make estimates in these seven regions more precise. Yet the concern here is compare residents of treated municipalities (in the seven regions) to residents in untreated municipalities (in other regions). As discussed in the memo below on retrospective evaluation, it may be useful to compare respondents in treated municipalities to respondents in a subset of untreated municipalities, say, those just across the border from treated municipalities. In a nationwide probability sample, estimates of quantities in the latter group may be most imprecise, due to small sample sizes. An over-sample of residents in appropriate control municipalities might therefore be most useful for purposes of estimating program impact.

Absence of randomization. The most appropriate way to select control units would be by random assignment of municipalities – *inside* the seven non-randomly selected regions – to receive interventions. We discuss this issue further below. Suffice it to say here that municipalities were *not* randomly selected for treatment (nor were the seven regions in which USAID/Peru focused its support of decentralization efforts).

One other point that may be useful in passing is that pilot districts were also non-randomly selected. As USAID's Request for Proposal (RFP) remarked in discussing the planned rollout of the program, "The phasing was designed to take into account the need to show early program success, and therefore includes some municipalities which already have been classified…as having high management capacity and a high level of resources" (USAID/Peru 2002). As I will discuss below, it may have been worthwhile to consider the possibility of random assignment of pilot districts. Stratified sampling could have been used to accomplish the goal of including municipalities with high management capacity and a high level of resources.

Actual cost of monitoring and evaluation

It is difficult to estimate the actual costs of monitoring and evaluation with any degree of precision. Monitoring and evaluation activities were not given separate line-items in program budgets. There are also disparate activities that fall under the rubric of monitoring and evaluation activities, and including all of them in a single estimate is challenging.

However, back-of-the-envelope calculations suggest that the costs of monitoring and evaluation under the existing program were substantial. According the contractor, the quarterly indicators gathered by local sub-contractors constituted the biggest single cost.[18] Some of these indicators were required for USAID reporting purposes, while others were gathered for internal performance-monitoring purposes.

[18] Interview, Tom Reilly, ARD; Lima, Peru, June 22, 2007.

According to local sub-contractors interviewed in Ayacucho, gathering and reporting these output indicators took around 50 percent of their time in 2004, and 20 percent in 2005, 2006, and 2007. (The percentage of time allocated to these activities decreased both because of changes in the internal reporting technology and because team members became more adept at gathering the information). Gathering the quarterly indicators associated with monitoring and evaluation therefore represented a substantial allocation of time.[19]

These local sub-contractors constitute one group of project employees in just one department. For purposes of estimating the overall cost of staff time involved in gathering these indicators, we need:

1) To determine whether this is representative of the percentage of staff time spent on these activities by other sub-contractors, both elsewhere in the department of Ayacucho and in other states, perhaps by sampling other sub-contractors;

2) To obtain estimates from the contractor of per-hour salaries and total hours worked.

The USAID/Peru contractor on the decentralization project has offered to help us estimate these costs, but I have not yet successfully coordinated with the contractor on this.[20] However, it is likely that these monitoring and evaluation activities represent a non-trivial proportion of the overall budget of $17 million spent over five years ($20 million were budgeted for the program).

Best-possible designs

In this sub-section, I discuss best-possible designs from the perspective of program evaluation. First, I discuss what an ideal ex-ante design for the decentralization program might have been in 2002, when the program was begun. Second, I also discuss how an experimental design might be employed in a second phase of the program, given that all the municipalities in the seven regions were already treated in the first phase. Since the USAID/Peru Mission is currently planning a second five-year phase of implementation, it is hoped that these comments might be useful to Mission staff.

A "tabula rasa" design

I assume that the decentralization program will be implemented in the seven non-randomly chosen regions in which USAID commonly works; inferences about the effect of the intervention will then be made to the districts and provinces that comprise these regions.

[19] Interview with the CEISA team; Ayacucho, Peru, June 27, 2007.

[20] The director of the PRODES program is currently on vacation; we may be able to ascertain these costs with more certainty when he returns.

The simplest design would involve randomization of treatment at the district level. Districts in the treatment group would be invited to receive the full bundle of interventions associated with the decentralization program (e.g., training in participatory budgeting, assistance for civil society groups, and so on); control districts would receive no interventions.

There are two disadvantages to randomizing at the district level, however. One is that some of the relevant interventions in fact take place at the provincial level.[21] Another is that district mayors and other actors may more easily become aware of treatments in neighboring districts (this issue is discussed further below). For both of these reasons, it may be useful to randomize instead at the provincial level. Then, all districts in a province that is randomly selected for treatment would be invited to receive the bundle of interventions.

Several different kinds of outcome measures can be gathered. Survey evidence on citizens' perceptions of local government responsiveness will be useful; so may be evaluations of municipal governance capacity taken across all municipalities in the seven regions (both treated and untreated).

A difference in average outcomes across groups at the end of the program – for example, differences in the percentage of residents who say government services are "good" or "very good," or the percentage who say the government responds "almost always" or "on the majority of occasions" to what the people want – can then be reliably attributed to the effect of the bundle of interventions, if the difference is bigger than might reasonably arise by chance.[22]

One feature of this design that may be perceived as a disadvantage is the fact that treated municipalities are subject to a bundle of interventions; thus, if we observe a difference across treated and untreated groups, we may not know which particular intervention was responsible (or most responsible) for the difference. Did training in participatory budgeting matter most? Assistance to civil society groups? Or some other aspect of the bundle of interventions?

This problem arises as well in some medical trials and other experiments involving complex treatments, where it may not be clear exactly what aspect of treatment is responsible for differences in average outcomes across treatment and control groups.

It seems preferable at this stage to design an evaluation plan that would allow USAID to know with some confidence whether a program financed by USAID make any difference. Bundling the interventions may provide the best chance to estimate a causal effect of treatment.

---

[21] Some interventions also occurred at the regional level, particularly towards the end of the program, yet these interventions constitute a relatively minor part of the program.

[22] Standard errors may need to be adjusted to account for the clustering of treated districts within provinces.

Once this question is answered, one might then want to ask what aspect of the bundle of interventions made a difference, using further experimental designs. However, another possibility discussed below is to implement a more complex design in which different municipalities would be randomized to *different* bundles of interventions.

The intention-to-treat principle can be used to analyze the results of the experiment. Some municipalities assigned to treatment may refuse to sign participation agreements or otherwise may not cooperate with the local contractor; these municipalities may be akin to non-compliers in a medical trial. In this context, estimating the "effect of treatment on the treated" may be of interest (see the Appendix).

It may be worth choosing pilot districts at random as well. In the first phase of the implemented decentralization program, only 145 municipalities were incorporated in the program in the first year, out of 536 that were eventually incorporated. USAID/Peru documents, as mentioned above, suggested that the phasing took into account "the need to show early program success, and therefore includes some municipalities which already have been classified…as having high management capacity and a high level of resources" (USAID/Peru 2002). Obviously, comparing municipal capacity across incorporated and unincorporated municipalities at the end of the pilot period may be virtually meaningless; the incorporated municipalities were *chosen* for their high degree of capacity. It would be much more meaningful to randomly assign municipalities for inclusion in the pilot phase. To the extent it is necessary to include some municipalities with high ex-ante management capacity and resources, this may be accomplished through stratified sampling of municipalities.

Second-phase design

USAID/Peru is preparing to roll-out a second five-year phase of the decentralization program, possibly again in the seven regions in which it typically works. At this point, all municipalities in the seven regions were already treated (or at least targeted for treatment) in the first phase. This may raise some special considerations for the second-phase design.

Our understanding is that there are several possibilities for the actual implementation of the second phase of the program; which option is chosen will depend on the available budget and other factors.

One is that all 536 municipalities are again targeted for treatment. As in the first-phase design, this would not allow the possibility to partition municipalities in the seven regions into a treatment group and controls.

In this case, the best option for an experimental design may be to randomly assign different treatments -- bundles of interventions – to different municipalities. While such an approach will not allow us to compare treated and untreated cases, it will allow us to assess the relative effects of different bundles of interventions.

This may be quite useful, particularly for assessing the question raised above about which *aspect* of a given bundle of interventions has the most impact on outcomes. Do workshops on participatory budgeting matter more than training civil-society organization? Randomly assigning workshops to some municipalities and training to others would allow us to find out.

A second possibility for the second phase of the program is to reduce the number of municipalities treated, for budgetary reasons. Suppose the number of municipalities were to be reduced by half. The best option in this case is probably to randomize the control municipalities out of treatment, leaving half of the universe assigned to treatment and the other half in control. Those municipalities assigned to treatment would be offered the full menu of interventions in the decentralization program.

Of course, randomizing some municipalities out of treatment is sure to encounter displeasure among authorities in control municipalities. Yet if the budget only allows for 268 municipalities assigned to treatment and 268 to control, this displeasure will arise whether or not the allocation of continued treatment is randomized. In fact, as discussed below, it may be that using a lottery to determine which municipalities are invited to stay in the program is perceived as the fairest method of allocating scarce resources.[23]

Finally, a third possibility could involve randomized roll-out of the treatment to different municipalities. This may be useful if the roll-out is not concluded by the second year of the program, as it was in the first-phase intervention.

Cost of evaluation under the best-possible designs

The need to gather outcome measures on control units – both through surveys of residents in untreated municipalities and through independent evaluations of municipal capacity in control districts – will mean an additional cost of program evaluation.

However, it is worth bearing in mind that such additional costs would have likely represented only a small fraction of the cost of the overall program as well as of the portion of overall costs going to evaluation. For example, adding 500 respondents from appropriately chosen control municipalities would likely have cost no more than $10,000, a drop in the bucket compared to the overall program budget of $20 million over five years.[24]

In addition, with appropriate design modifications, there might have been substantial net savings. One possibility for cost-savings would involve substantially limiting the volume of output/outcome indicators gathered by each of the local sub-contractors. The USAID/Peru contractor made the very useful suggestion that measures could have been *sampled* across local jurisdictions, rather than gathered quarterly on the entire universe of

[23] For reasons discussed above, it may also be useful to conduct the randomization at the provincial rather than district level.

[24] This assumes a cost of $20 per respondent; according to Patricia Zárate, IEP, costs per respondent can vary between $14 and $20.

536 municipalities. A related idea is that local sub-contractors could have been told that they would have to go gather the indicators and report on them each quarter with some positive probability; but they would not actually have to do so in each quarter.

Given the estimate of local sub-contractors in Ayacucho that 50 percent of staff time was spent on indicator reporting in the first year of the program, and 20 percent thereafter, probabilistic auditing could have represented a substantial savings. Suppose that data reporting was required of local sub-contractors with probability 0.25 each quarter, so that on average sub-contractors gathered and reported indicators once a year. This might imply costs in staff time of one-quarter what they actual were. If $4 million were previously spent on staff time for data reporting, $1 million might be spent under the new scheme.[25]

2.2.2 Other examples: Rule of law, political parties, and extractive industries

Several of the programs planned under the new strategic assessment might also be amenable to randomized designs. In this section, I briefly review possibilities for experimental designs afforded by programs related to the rule-of-law, political parties, and extractive industries. I also include a text box on rural connectivity initiatives.

Rule of law (the IRIS project)*:* Most of the interventions under the rule-of-law programs implemented by IRIS (a project described further below under "non-experimental designs") were not amenable to randomization across units. However, there were one or two interventions that could in principle have been randomized.

For example, after the passage of a new penal code, some judges in district courts were switched to the new system of judging cases while others were left to clear the backlog of cases that had already entered the courts under the old system. Under the observational (non-experimental) evaluation plan that was actually adopted, cases administered by judges under the new system were compared to cases administered under the old system. Comparisons were made across groups with respect to variables such as the average time to disposition of the court cases.

This non-experimental design represented a valuable (and rare) evaluation plan: there was an actual comparison made across treated and untreated units on an outcome measure of interest. In this and similar example, the data seemed to show a substantial effect of treatment.

However, judges were non-randomly assigned to stay in the old system or migrate to the new one (the chief judge apparently decided who would move). This raises the possibility that characteristics of judges who stayed or migrated are partially or wholly responsible for differences in the average time to disposition.[26]

[25] Of course, it is difficult to know from our small sample of programs how typical such savings would be.

[26] While data were not available, it would have been helpful to compare the *difference* in time to resolution, before and after the switch of systems, among judges who switched and judges who didn't; this

In principle, it would have been possible to assign district court judges the old and new systems at random. While the research design idea is straightforward, however, it was likely to be politically difficult: chief judges may not want to relinquish power over these assignments.

I do not comment further here on the political feasibility of randomization in this specific program. At a minimum, it seems that it would have been necessary to obtain the agreement of political actors at the top of the Justice Ministry to a randomization plan. This specific example may suggest the general importance of convincing political actors at the top of the useful (or at least non-threatening) nature of randomization. The next example suggests this point as well.

Political parties*:* One future idea under the new political parties program could be to provide assistance to the major national-level parties in opening or strengthening local offices in selected municipalities. If the parties themselves might choose where to open offices, however, the design would be non-experimental.

Moreover, if outcomes are not tracked in municipalities in which USAID partners do *not* support local party offices (i.e., controls), inferences may be especially misleading. Suppose measures are taken today and in five years of local party strength (see below) and an increase is found. Is this due to the effect of party strengthening activities supported by USAID? Perhaps. Yet it could be due to some other factor, like a change from an electoral system with preferential voting to closed party lists, which would tend to strengthen party discipline and, perhaps, local parties; such a change is currently being considered in Peru.[27] The point is that without data on controls, it will be impossible to separate the effect of USAID local activities from the effect of the law.

At a minimum, then, it would be advisable to consider gathering data on control municipalities. In addition, while an experimental approach may not be feasible in this instance, it is certainly possible in principle, and such an approach would provide a stronger basis for impact attribution than a non-experimental design.

Under an experimental design, USAID or the local implementer would select municipalities in which to establish or strengthen local parties randomly, from a set of acceptable municipalities. Local parties would have to accept that USAID or the contractor would select the municipalities. There may be ways to overcome any resistance to such a plan, however; for instance, a party such as Unidad Nacional (the

---

would have required pre- and post-switch data on both groups of judges. While still non-experimental, this comparison would lend greater confidence to the claim that the switch in systems had a causal effect on the time to resolution of court cases.

[27] In the current electoral system, there is proportional representation at the department level, and voters vote for party lists but can indicate which candidate on the list they prefer; according to a range of research on the topic, this can create incentives for candidates to cultivate personal reputations and also makes the party label less important to candidates. Under a closed-list system, voters simply vote for the party ticket, and party leaders may decide the order of candidates on the list. This may tend to increase party discipline and cohesion (as well as the internal power of party elites).

rightist party whose candidate in the 2001 and 2006 presidential elections was Lourdes Flores Nano) has almost no base outside Lima and might accept any help it can get to broaden that base. Another obstacle is that parties may want to target certain kinds of municipalities, for example, those where they already have some support. It may be helpful for this purpose to stratify municipalities -- for example, by past levels of electoral support for each party -- and conducting the randomization within strata.

Outcome indicators might include the municipal vote share of each party in subsequent elections, with comparisons being made across treated and untreated municipalities; there may be other, harder-to-measure outcomes of interest, too.

Inferences may be complicated if more than one party opens or strengthens an office in the same municipality (i.e., if there are two parties and both are strengthened locally, party vote shares may be unchanged. This concern may be lessened by the fragmentation of the party system and by the current local dominance of regional parties. In recent regional elections, for example, 23 different regional parties won office across Peru's 24 departments; these regional parties differ from the national parties whose local roots USAID seeks to strengthen.

Extractive Industries: There is currently a very small pilot program that seeks to promote dialogue in two mining communities between the State, companies and local citizens, with the larger goal of "decreasing the probability of social conflict."

This program has the advantage of possessing a relatively easy-to-measure outcome variable, social conflict (compared to, say, transparency). For example, this variable might conceivably be proxied by the annual number of local marches/demonstrations. However, without comparing mining communities with which USAID works to those with which it does not, it will be difficult to evaluate the causal impact of the program on decreasing the probability of social conflict.

In a future roll-out of the program, mining communities with which USAID might work could be randomly selected from the set of eligible mining communities. This would provide the most secure basis for attaching a causal interpretation to a finding that, e.g., there were fewer marches and demonstrations in communities in which USAID worked than in those in which it did not work.[28]

[28] This is not a typical Democracy and Governance program in cross-national perspective, but since it is managed by the DG office in Peru, I comment on it here.

Text box: Rural connectivity

In 2007, the government of Peru approached USAID/Peru for assistance with the rollout of community computer centers (also known as "telecenters") in 84 selected rural municipalities. The plan called for USAID to fund initial Internet service in the municipalities, all located in seven regions of the country in which USAID has ongoing programs.

Does the availability of local telecenters encourage broader public access to government services, greater citizen involvement in politics, or more favorable perceptions of government transparency? Studying the telecenter rollout in Peru could allow USAID to answer such questions about the impact of a program it supports.

The issue, however, is how best to design the program to allow for rigorous evaluation. In principle, this program would allow for a randomized design. In practice, program rollout has already begun, and municipalities have been invited to participate on a non-random basis.

I discuss in this text box how an experimental design could have been designed to evaluate the impact of this program. I then comment on how the effects of the program might be studied under the non-experimental design actually employed.

An experimental design

There are 537 municipalities in the seven regions in which USAID works; around 200 of these municipalities lack Internet access altogether and are therefore eligible to participate in the government's telecenter program. (The government requires that telecenters be located in municipalities that currently lack Internet service, so as to preclude competition with private providers). Under the terms of the program, the municipalities must provide some of the financing for the telecenters. Local authorities in selected communities must therefore agree to participate in the program.

In an experimental design, invitations to participate in the program would be issued at random to authorities in eligible municipalities. Municipalities receiving invitations would constitute the "treatment" group; eligible municipalities not selected for invitations would be the controls. For practical reasons I discuss below, rather than invite only 84 municipalities to participate, it might be useful to select perhaps 100 of the 200 eligible municipalities at random for invitations; the other 100 municipalities would serve as controls.

Measurements on outcomes of interest could then be gathered on both the treatment group and the controls at the end of a given time period (say, eighteen months after the start of the program, which is the intended length of USAID financing). Differences in average outcomes across the two groups could then be reliably attributed to the causal effect of the program, subject to the usual caveats.

The intention-to-treat principle, discussed above, may be useful in analyzing the results of the experiment. In some communities offered telecenters, authorities may not agree to participate in the program; others may do so but then never raise the necessary self-financing to open the centers, or centers may close before the end of the program. Since local characteristics that influence whether participation agreements are signed or telecenters remain open may also influence outcomes of interest, it can be misleading to compare communities that actually have telecenters at the end of the program to those that do not.

That is why the intention-to-treat principle provides a robust way to assess the causal effect of the USAID-financed telecenter roll-out. According to this principle, we would compare municipalities that are randomly invited to participate to those that are not invited -- rather than municipalities that actually received telecenters and those that did not. Confounding local characteristics should be independent of whether the community was *invited* to participate in the telecenter program, since invitations are issued randomly. (It may also be useful to consider the "effect of treatment on the treated;" see the Technical Appendix).

Use of the intention-to-treat principle carries some implications for program design. For example, since some communities may elect not to participate, it will be useful for USAID staff and government partners to estimate the refusal rate ex-ante and adjust the number of randomized invitations accordingly. If the expectation is that roughly 15 percent of communities will refuse, and there is funding available for 84 communities, it may make sense to issue invitations to 100 communities, and make comparisons between these 100 invited communities and the 100 uninvited municipalities (as recommended above). Of course, in this case more than 84 communities might end up accepting telecenters; if the budget really caps the number of municipalities at 84, then one might want to randomize fewer municipalities to treatment.

Perhaps the greatest practical issue would involve securing the consent of the government (here, the Ministry of Transport and Communications) to the randomization plan. However, the Foreign Service National (FSN) in charge of the program at the USAID mission in Lima did not think this would have posed a particular issue in this case. To the extent the government insisted that one or another municipality be included in the program, one workaround would be to guarantee a telecenter to this community, but leave it out of the set of communities that will be randomly assigned to treatment or control – and thus out of the set of municipalities that will provide the basis for the experimental comparison.

A quasi-experimental design

In point of fact, the communities will not be randomly selected; communities have begun to sign participation agreements at the time of writing.

What are the inferential issues? Suppose we compare communities that have Internet access via telecenters at the end of the program to those that do not and find that citizens in the former communities tend to use local government services at greater rates. Does this mean the telecenters had a causal effect? Perhaps not. It may be that local governments which were selected to receive telecenters and then actually accepted the centers were already more effective at providing services, and therefore citizen use of the services was already greater.

Nonetheless, there may be valuable opportunities for impact assessment given the existing design of the program. One key is to conduct both pre- and post-tests on outcome variables of interest in communities that receive invitations to participate and those that do not. Even though the treatment is not randomly assigned, it may be worth making comparisons across these "intention-to-treat" groups, i.e. across the invited groups, rather than across municipalities that do and do not have telecenters at the end of the program; the intention to treat may be subject to fewer inferential issues than receipt of treatment. (Further discussion is in the memo on non-experimental design).

2.3 Standard objections to randomized designs

In this section, I outline several objections to experimental designs that were raised by field interviewees, including USAID staff and local implementers. By calling these "standard" objections I do not intend to be dismissive of them: many are very important and sometimes can constitute real obstacles to using experimental designs.

However, there are often workarounds that can help to partially or wholly counter such objections. The point of this section is to comment on some potential workarounds, inspired by our field visit to Peru, that may be more generally useful.

Political will

> Perhaps the most common objection raised by USAID staff and its contractors concerned the importance of "political will." USAID and its contractors can often only work with local authorities that accept their help. The temptation, moreover, may be to work with local authorities that seem most disposed to work with USAID, since working with these local authorities seems to promise the greatest potential for "impact."
>
> In an experimental design, one workaround is to select the set of municipalities or other units that are eligible for treatment, on the basis of political and other criteria. Randomization to treatment and control should occur within this group of eligible units, and only those units randomly assigned to treatment should be invited to participate in the program.
>
> Some invited municipalities may not participate, due to an absence of political will. These municipalities are like non-compliers in a medical trial. With enough municipalities, non-compliers will be balanced across the treatment and control groups, and USAID contractors may be assured there will be some units with political will, among the municipalities invited to participate.
>
> The intention-to-treat principle should then be used to analyze the results of the experiment; it may also be useful to consider the effect of treatment on the treated (see the Appendix).

Some units must be treated, for political or other reasons

> For political or other reasons, allocating treatment to one or several units may be non-negotiable. Randomization to treatment and control should then occur among eligible units who do not have to be treated for these reasons. The key is that outcomes should be compared across the randomly-assigned units, and the non-randomly selected units that had to be treated for political reasons should be left out of the comparison.

Of course, one can always look as well at outcomes in the non-randomly selected units. Yet comparing outcomes in such units to non-treated 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.

Ethical implications of denying treatment to controls

Is it ethical to deny treatment to control groups? This issue comes up in many medical trials as well. The standard defense may be relevant in the present context as well: without an experiment, how do we know treatment helps? USAID intervenes to assist democracy and governance projects all over the world; just as in the medical field, it behooves us to know with as much confidence as possible what works and what does not.

Perhaps a more palatable defense of experiments, from the point of view of units that might be assigned to either treatment or control, is that there will virtually always be untreated units (at least in the sorts of programs that are amenable to experimental designs). In the context of a decentralization program, it is infeasible for USAID to work with all municipalities. The only question is how untreated units will be chosen.

As discussed below, in many contexts it may be fairest, and most ethically defensible, to choose untreated units by lottery.

"Jealousies" of untreated units

How can implementers manage political issues arising from the "jealousy" of non-treated control units? Several suggestions might help with this issue.

First, the experimental design itself may help decrease the probability that authorities in control units become aware of treatments administered to neighboring, treated units. In a decentralization program, for instance, randomization can take place at the regional or sub-regional (provincial) level: all municipalities in a given region or province then receive the treatment or the control.[29]

This approach may not help entirely, however. In Ayacucho, for instance, we found that provincial mayors sometimes gather in regional capitals, either to participate in USAID-financed trainings or for collective meetings with regional authorities. If randomization takes place at the sub-regional (provincial) level, such contacts could provide opportunities for mayors of control provinces to learn about interventions in treated provinces.

[29] Statistical analysis of experimental data can take into account the consequent clustering, for purposes of variance calculations.

Second, some experimental designs involve not an absence of treatment in control units but rather the implementation of a different treatment. In USAID/Peru's decentralization program, as we discussed above, 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 sub-regional level. Though inter-municipal and inter-provincial meetings might again 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 ethnical implications of randomization. There are situations in which it may seem *fairer* to use a lottery to randomize units out of treatment. The second phase of USAID/Peru's decentralization program may involve fewer municipalities than the first phase, due to budget constraints. Is it not most politically palatable to tell municipalities no longer receiving assistance that the municipalities were chosen for the second phase by lottery?

Difficulties in mitigating "jealousies" of control units might also be bolstered by the fact that in Democracy and Governance programs, USAID is often not providing material assistance but rather logistical and technical support.

Putting out fires

As was noted to us by a political officer at the US Embassy, the Embassy sometimes is compelled to "put out fires." In an experimental evaluation of the impact of municipal-level interventions in mining towns on the likelihood of company-community conflict, the Embassy may have to intervene when conflict breaks out in a community – whether that community is in the treatment group or in the controls.

This may or may not pose an issue for experimental inference. Some "fires" may be independent of treatment assignment, that is, they may be as likely to occur in treated and control units. Other "fires" may reflect, say, the absence of treatment among controls; this may raise more serious issues.

Other donors flood the controls

Another issue relates to donor coordination. Although we received conflicting reports on this score, it is possible that donors from other countries may concentrate their programs in areas in which USAID does not work. One program officer suggested that other donors might "flood the controls" in an experimental design (since USAID is not there).

This may be a real issue for inference. If donors from other countries explicitly go control units with which USAID is not working, there is a confounding factor

associated with treatment assignment; this could certainly bias inferences about the effect of USAID interventions.

However, it might be pointed out that if anything, this is likely to dilute the (hopefully positive) effect of treatment. If other donors flood the controls and we still see a difference between groups, we can infer a causal effect of USAID's intervention. (At least, we can evaluate the effect of USAID relative to other donors!).

A possible response to this issue is not to advertise the existence of control units, say in the context of a decentralization program; it is known that USAID works in seven regions, but it need not be known which municipalities it is working with in the context of a specific program. Another is to randomize different treatments across all municipalities (as was suggested above for the second phase of the decentralization program). In other words, USAID would work with *all* municipalities in the seven regions, obviating the concern about donors flooding the controls; yet USAID could randomly assign different treatments to different municipalities.

Contamination

Discussed above; this is an inferential issue, however, not a political one. In standard models of experimental inference, the response of one unit to treatment is not affected by the response of other units.[30] Violations of this assumption can raise problematic issues for inference.

This is likely to be an issue in some experimental evaluations of USAID programs. Sometimes, design modifications can help. For example, in the context of the decentralization program, randomizing at the provincial level might decrease the probability that district mayors are aware of treatments administered to other units.

Gathering outcome data on controls

One Foreign Service National (FSN) working on monitoring and evaluation at USAID/Peru Mission suggested that it might seem "morally incorrect" to gather data on control groups, say, in the context of the decentralization program. Some indicators will be more intrusive and difficult to gather on control units than others.

Ethics aside, this raises a different but important point: a lot of USAID data on treated units are gathered in the context of ongoing relationships between contractors and local authorities, and gathering such data on control units may be

[30] In Rubin's (1974) formulation of the standard model of experimental inference proposed by Jersey Neyman (1923; see Dabrowska and Speed 1990), this is called the "stable unit treatment value assumption" (SUTVA); see also Holland (1986).

much more difficult. This is obviously an issue that may be relevant for non-experimental designs as well.

However, much of the difficulty of gathering data on controls probably relates more to *output* measures than 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. However, coding outcome measures, such as survey responses on government responsiveness or evaluations of municipal capacity, may raise no different issues among the controls as the treated units.

Heterogeneity of treatment effects

Some program officers raised the issue of heterogeneity of treatment effects. For instance, in a decentralization program, interventions might have a big impact in some localities and a negligible impact in others.

There is tremendous heterogeneity across the municipalities in Peru. (47 districts have one inhabitant per kilometer and another 22 have 10,000 residents per kilometer). It is reasonable to think this might imply heterogeneity in the effect of treatments as well. Yet this is generally the case with experiments, which always help us estimate the average response to treatment.

Randomization among sub-groups would help us estimate the impact in distinct kinds of municipalities. For instance, one could randomize treatment within more densely-population urban districts and within more sparsely-populated rural districts.

3. Non-experimental designs

While experiments can offer a feasible and cost-effective evaluation technique in some instances, they will not in others. It is therefore important to consider non-experimental designs, which I do in this memo.

For some USAID programs, good non-experimental designs should share some of the same basic features as experimental designs. First, it is crucial in nearly all programs measure outcomes, rather than use indicators that only monitor contractor performance. Second, in programs where a relatively large number of units are treated, it is also essential to gather data on control units.[31] These data should ideally be measured (at a minimum) before the intervention and at the end of the program; this can increase our ex-post confidence in inferences.

Consider a non-experimental design for the political parties program discussed above, in which USAID partners will work with national political parties in selected municipalities to strengthen the parties' local offices. One possibility is gather baseline data (say, local party vote share in the previous election) and data towards the end of the program (local party vote share in the subsequent election).[32] The observation that party vote share has increased, on average, in the municipalities in which USAID worked may be due to the effect of the program. Yet as discussed in the previous section, this inference is vulnerable to the charge that all municipalities were affected by some national-level development, such as a new law getting rid of the preferential vote.

By contrast, pre- and post-intervention data on control municipalities would give us greater confidence in the inference that USAID had an impact: we would compare the difference in local party vote share in the treated group to the difference in local party vote share in the controls.

Of course, the non-random selection of treated units can still present inferential issues: treated municipalities might have been selected because they are expected to be more responsive to treatment, so the average outcome in the treated group doesn't give us a good estimate of what would have happened if USAID had instead treated the control group. Yet comparing pre- and post-intervention data on treated and untreated units can offer the soundest basis for causal inference in many non-experimental designs.

In some USAID programs, however, it will not be possible to gather data on controls, perhaps because there is only one unit to be treated (or not).[33] In such cases, other kinds of evidence might be useful. There is a long social-scientific tradition, for example, that

[31] How do we know when the number of units is large enough? In some contexts, explicit ex-ante calculations of statistical power may be useful for deciding whether it is worth it to collect data on controls.

[32] This may not be the most valid or useful outcome indicator, particularly if USAID is working with more than one national party in given municipalities; see the discussion in the previous memo.

[33] It may be apocryphal, but one expert on medical trials apparently always recommended randomizing the first patient (!). Joking aside, the point is that there's no way to use another unit to get an estimate of what would happened if the treated unit had not been treated (or if the untreated unit had been treated).

emphasizes the use of counterfactual reasoning in case studies (see Fearon 1991), as well as modes of causal inference in which key nuggets of information – causal "smoking guns" or "causal process observations" (Collier, Brady and Seawright 2004) – play a key role.

In some of the programs we examined, evidence in this mode appeared to provide the best, and perhaps only, available way to evaluate the impact of the program. My comments will be relatively brief on this topic, but I offer two examples drawn from our discussions in the field, in the hope that these will be useful to members of the committee.

3.1 Case studies: Transparency and legislative strengthening

Transparencia. The director of Tranparencia, a local NGO that seeks to promote transparent and inclusive institutions and the formation of a politically-active and vigilant citizenry, underscored the important role of the organization in promoting a new law governing political parties.

How did the director know that Transparencia had an impact? First, leaders of the parties met at the Transparencia office to negotiate approval of the law. Second, Tranparencia itself played an important role in writing sections of the draft legislation. Third, "everybody knows" that Transparencia played a crucial role.

On the one hand, these pieces of evidence – especially the claim that Transparencia organized the meeting at which the draft law was negotiated by party leaders -- might suggest a "smoking gun" or "causal process observation" (Collier et al. 2004) that allows one to infer that Transparencia had a causal impact on the passage of the law.

On the other hand, it is clearly difficult to answer perhaps the key counterfactual question: would the law have been passed if Transparencia did not exist (or if it had stayed out of negotiations over the bill)? It is hard to know with confidence, though the evidence may be suggestive.[34]

SUNY. USAID/Peru funded a large legislative strengthening project, in partnership with the Center for International Development of the State University of New York (SUNY). Goals of the program included increasing the speed and efficiency with which the legislative agenda was initiated and considered, as measured for example by the percentage of initiated legislation that was considered during the annual session; and promoting the transparency of the legislative process, as measured for example by the percentage of total legislation that included an organized public hearing (SUNY 2004).

In one sense, there was only one unit treated in this program: the Peruvian Congress. Data on indicators such as those mentioned in the previous paragraph were gathered on a

[34] Indeed, some people with close ties to the parties did not attribute that influence to Transparencia, in conversations with members of our team.

semi-annual basis over the life of the program, including baseline data at the start of the program (2002-2004).

The data allow the construction of an interrupted time-series, where legislative outcomes in the Peruvian Congress after the start of the program can be compared to outcomes prior to the program. However, we can only estimate the counterfactual legislative efficiency -- what legislative efficiency of the Congress would have been if the intervention not been applied -- by looking at the value of the pre-intervention (baseline) indicators.

This strategy may not provide the most robust basis for inference, since many other factors that might affect legislative efficiency may have been changing over this period (e.g., some analysts characterize this as a period of democratic "retransition" after the resignation of Alberto Fujimori and election of Alejandro Toledo as president).

Thus, other kinds of strategies for inference – for instance, finding causal process observations that suggest the impact of the program implemented by SUNY local partners -- may be useful. Interviews with involved actors could be a source of such information. Of course, this approach may not fall far from the anecdotal, and it seems difficult as a general matter to make recommendations as to how causal process observations should be found.

4. Retrospective Evaluations

One question USAID has asked the National Academy of Sciences committee concerns the most useful ways to conduct retrospective analysis and program evaluation. This memo considers this question based on analysis during our field visit to Peru.

4.1 Natural experiments

In principle, one possibility is to seek to find and exploit "natural experiments" that would be useful for evaluating the impact of USAID interventions. Natural experiments are situations in which natural – here, social and political – forces conspire in a way that one can credibly argue that assignment to treatment is "as if" random.

A famous example of a natural experiment comes from epidemiological studies of cholera transmission by the 19th century anesthesiologist John Snow. At mid-century, two companies supplied water to many areas of London; the companies' intertwined pipes served water to side-by-side houses and to residents rich and poor, young and old. A few years before a major cholera outbreak, one of the water companies moved its intake pipe upstream on the Thames, to a purer source of water, while the other left its intake pipe in place. This move of the water pipe very likely took place without the consent or knowledge of most residents. Snow compared cholera death rates in houses served by both companies and showed dramatically lower death rates in houses served by the company that had moved its pipes. Since the source of water supply was credibly "as if" random with respect to other factors that might cause death from cholera, Snow's study provided convincing evidence that cholera could be transmitted through water (Snow 1855; see Freedman 1991, also Dunning forthcoming).

In some USAID programs, like the decentralization program described above, some units (departments, municipalities, etc.) have received a USAID intervention and others have not. In order to pass as a natural experiment, however, the assignment of the intervention to units would have to be plausibly "as if" random and, in particular, not correlated with characteristics of the units that might account for differential outcomes across the treatment and control units.

The possibility of exploiting such natural experiments appears limited, at least in the Peruvian context. Units are chosen non-randomly to receive interventions. In most cases, treatment assignment does not appear independent of unit characteristics that would affect outcomes.[35]

For example, USAID/Peru has chosen to work in seven departments of the country because those departments share indigenous Andean and Amazonian populations, high poverty rates, narcotics activities, and a history of violence (stemming from the conflict with the Sendero Luminoso). This purposive choice is no doubt valid for strategic reasons, but it complicates the task of retrospective program evaluation.

[35] However, see the discussion below of the possibility of matching municipalities across borders.

In the context of the decentralization program described above, for example, one might want to compare municipal-level outcomes in the 536 treated districts (inside the seven departments) to outcomes in districts outside the region. The problem is obviously that any differences could be due to the initial disparities between municipalities outside the regions and the poor, violence-ridden municipalities inside them.

One standard approach would be to compare the over-time difference in outcomes in untreated municipalities outside the regions to treated municipalities inside the regions. This would pre-intervention as well as post-intervention measures on treated and untreated municipalities, and the former in particular may be very difficult or costly to obtain.

Difficulties in collecting pre-intervention measures aside, however, this "difference-in-differences" approach may not help much for discerning the causal impact of treatment. For instance, differential outcomes over time – say, faster "growth" of municipal capacity in the treated regions -- could well be due to recent bounce-back from the worst years of the war, not to the effect of treatment.

In sum, it seems rare that past USAID interventions could be characterized as "as if" randomly assigned; we did not find credible examples of "as if" random assignment in Peru. Many USAID programs are probably assigned in response to very specific developments on the ground: say, there has been a crisis in and weakening of the political party system, so a political parties program is designed. While valid for strategic purposes, this complicates the task of retrospective evaluation.

4.2 Controlled comparisons

There are, however, some useful possibilities for retrospective evaluations. Such studies will be weaker for purposes of causal inference than natural experiments but can nonetheless be informative.

Probably the best approach in the context of some USAID programs would be to pick a *subset* of untreated units to compare with the (non-randomly 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.

In the case of the decentralization program, for instance, a possibility is to compare districts just across some of the regional borders from treated municipalities. There are districts in the department of Huancavelica, for example, that might provide good comparison groups to the treated municipalities. Like the seven regions in which USAID works, Huancavelica is a very poor department with a large Andean indigenous population, and it was a site of violence during the Sendero Luminoso years. However, it was not selected as a region in which USAID would work.

It is likely, then, that some municipalities in Huancavelica credibly shared similar initial conditions as treated municipalities, at the start of the decentralization program. Such untreated municipalities might also be similarly subject to "bounce-back" from the effects of the war, like the treated municipalities, potentially eliminating one rival explanation for any differences across the groups.

Indeed, when municipalities on either side of the borders are carefully selected for retrospective analysis, the comparison might even provide a credible natural experiment. This approach has provided the basis for natural experiments in various social-scientific studies (e.g. Posner 2004).

4.3 Outcome data on controls

A possible Achilles Heel of a retrospective evaluation plan, however, is again the lack of outcome data on control units. In the context of decentralization, data on municipal capacity in the appropriate untreated districts of Huancavelica may be slim.

Survey data may be exploited to tap residents' views of local government responsiveness in these districts (for example, the LAPOP survey). However, there may not be many survey respondents from these districts, which will imply lack of precision for estimates drawn from the survey data.

If USAID is serious about retrospective evaluations of program impact, it may be necessary to invest resources in identifying and gathering data on appropriate control units. In the decentralization case, it will likely be difficult to obtain good pre-intervention data on appropriate districts in Huancavelica, though perhaps not impossible.

However, it will be much more feasible to commission individual-level surveys or studies of municipal capacity in the districts. This may be very worthwhile to pursue further.

In sum, the recommendation for retrospective analysis is to concentrate resources in gathering appropriate data to be able to compare across treated and (appropriately chosen) untreated units. Concentrating resources in a few retrospective evaluations but carrying them out in the best-possible way is likely to create the most opportunities for learning retrospectively about the impact of USAID interventions.

5. Conclusion

The foregoing memos provide some suggestions for improvements to USAID program evaluation practices, drawing on experiences during a recent field visit to Peru. The main recommendations are to gather data on outcomes, include control units where possible, and introduce experimental (randomized) designs when feasible.

These may sound like fairly simple innovations, but in practice they would constitute a major shift in how USAID does program evaluation, at least in the programs we encountered in the field. Such improvements in program evaluation practices would allow USAID to know with greater confidence whether programs and interventions are having any impact, and what that impact might be.

As discussed in the overview memo, most of the advances would come at the level of "intermediate" hypotheses, however. For example, while the design modifications recommended here would not tell us whether decentralization in Peru has helped to stabilize democracy, they can tell us whether USAID programs have helped increase the responsiveness of local governments to citizens' demands – which is presumed to be an important step in stabilizing what the committee has called "big D" democracy.

This focus inevitably raises the issue of aggregating up to the bigger outcomes, though. Will we simply have to presume that decentralization stabilizes democracy? Can we ever accurately estimate the causal impact of USAID programs on such aggregate national-level outcomes?

Improving evaluation of meso-level hypotheses is certainly necessary for any eventual aggregation, and my hunch is that it offers the most promising entry point for improvement at this point. In addition, having solid knowledge of the impact of different USAID programs can eventually lend important points of leverage to the effort to evaluate the aggregate hypotheses.

Nonetheless, it is probably most realistic to suggest that there will continue to be important difficulties in aggregating up to convincing answers about the impact of USAID assistance on national-level democracy. Assessing hypotheses about aggregate outcomes using within-country evidence is not impossible: the use of counterfactual reasoning in case studies (see Fearon 1991) and modes of causal inference in which key nuggets of information can play an important role (Collier et al. 2004).

However, it is also probably the case that to the extent such hypotheses are tested, the evidence may often involve comparative (cross-national) data. Important inferential problems may arise in that context as well. For one, since USAID's Democracy and Governance (DG) assistance is non-randomly assigned across countries, funding allocation decisions may take into account characteristics of countries that are not or cannot be measured systematically, perhaps across a set of countries with similar measured characteristics (such as national income). These unmeasured/unobserved factors may in turn make some countries more or less likely to strengthen as democracies;

for example, DG assistance may be targeted to the countries that in the judgment of program officers and USAID staff have the "best chance" of democratic transition or consolidation. Even in the highest-quality studies, such as Finkel et al. (forthcoming), non-random selection and other issues can complicate estimation of the causal effect of USAID assistance.

This should not be read as a pessimistic but rather realistic conclusion about the possibilities for testing the impact of USAID programs at the aggregate level. USAID asked the National Academy of Sciences committee to recommend design modifications that would improve the evaluation of program impact, and the recommendations from the various field teams seem sure to help. Yet social scientists have struggled to understand the impact of development assistance and other factors on democracy for many decades, and though there are successes, there are also serious limitations to what we know.

Appropriate design modifications by USAID, including the introduction of experimental evaluation techniques, could constitute an important opportunity for improving our knowledge in these areas, however. Only by improving program evaluation can the impact of USAID programs be properly assessed and documented. Creating more rigorous knowledge of program impact through ex-ante design modifications is the best way to do this.

7.1 Appendix

In this appendix, I discuss further the "intention-to-treat" principle and the "effect of treatment on the treated." To do so, I draw on the rural connectivity initiative, discussed in the text box above, and develop an example with hypothetical data.

Suppose that we have 200 municipalities, 100 of which will be assigned at random to receive invitations to host telecenters in their communities. These 100 communities constitute the "intention-to-treat" group; the other 100 comprise the controls. Suppose that among the intention-to-treat group, authorities in only 85 municipalities accept the invitation and successfully keep telecenters open during the life of the program. This is the group of of municipalities that "accepted" treatment.

At the end of the program, we will gather outcome data in each community through surveys. We will focus for illustrative purposes on one candidate indicator, the percentage of the population that has knowledge of some local government service. For simplicity, assume that we sample 10 people in each municipality at random, and that each resident living in any of the 200 communities has an equal probability of selection into the sample.[36]

To estimate the effect of telecenters on citizens' knowledge, we should *not* compare the 850 respondents living in municipalities that accepted telecenters to the 150 + 1000 = 1,150 respondents who lived in municipalities without telecenters (see Table 1). This comparison can be misleading, because communities with more active governments and a more knowledgeable citizenry may be more likely to accept telecenters to start. Ex-post differences between these groups may therefore reflect pre-existing factors rather than the effect of telecenters.

7.1 Intention-to-treat analysis

Two better estimators of the causal effect of telecenters are available. The first is based on the intention-to-treat principle. According to the hypothetical data in Table 1, 400 citizens in the intention-to-treat group had knowledge of government services (first row, final column), while 250 citizens did among the controls (second row, final column).

[36] Equal probabilities of selection could result if (unrealistically) all municipalities were the same size and we conducted a simple random sample. If municipalities are different sizes, we must over-sample more populated cities to maintain a constant probability of selection across the universe of potential respondents; then we would rescale estimates, as suggested in the footnote below. Note that standard errors may need to be adjusted to account for clustering within municipalities.

We can therefore estimate that the random assignment to treatment - the "intention to treat" - raised individual responses by

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

that is, an estimated 60 percent.[37]

Table 1. The effect of rural telecenters: hypothetical data

	Group size (municipalities)	Group size (sampled individuals)	Number of respondents with knowledge of government services
Intention-to-treat group	100	1,000	400
Municipality accepted telecenter	85	850	370
Municipality refused telecenter	15	150	30
Controls	100	1,000	250
Municipality would have accepted telecenter (estimate)	85	850	?
Municipality would have refused telecenter (estimate)	15	150	30

In some sense, intention-to-treat analysis estimates the effect of what "we" do – that is, randomly assign municipalities to be invited for treatment – rather than what "they" do – accept a telecenter. Of course, the effect must presumably work through some municipalities actually accepting treatment; and it is a robust method of inference, because what we do is randomized and therefore independent of other factors that affect the outcome of interest.

Note, however, that the intention-to-treat analysis may result in a dilution of the treatment effect: it ignores the fact that some citizens in the intention-to-treat group live in municipalities that refused treatment, so telecenters could not have had a causal impact on their level of knowledge.

[37] If the groups were not of equal size, we would have to rescale the estimate; i.e., here we are calculating (350/1000)/(250/1000).

7.2 The effect of treatment on the treated

There may therefore be a better alternative in this context, which is to estimate the "effect of treatment on the treated." This estimate compares citizens in municipalities that accepted telecenters, among the intention-to-treat group, to citizens in municipalities that *would* have accepted telecenters had they been offered, among the controls.

The problem is that we can't immediately distinguish citizens in the latter group from other respondents among the control group, because we don't know which cities would have refused treatment had it been offered: that is why there is a question mark in the second line of the second row and final column of Table 1.

The workaround is as follows. First, note that 15 of the municipalities in the intention-to-treat group refused telecenters (third line, first row, first column). This is also our best estimate of the number of municipalities among the controls that *would* have refused the treatment, had it been offered. The randomization plays a key role: factors that cause communities to refuse telecenters should be balanced, on average, across the intention-to-treat and control group, because of random assignment.

Given the equal sizes of the intention-to-treat and control groups, we can also estimate that 150 of the citizens in the control group live in communities that would have refused treatment. Finally, we estimate that 30 citizens among the controls who reported knowledge of the government service live in such refusing municipalities. Again, we do not know which of the citizens in the control group live in such municipalities, but we can use the intention-to-treat group to estimate the number.

Since we know that 250 respondents in the control group reported knowledge of government services, and 30 of these lived in municipalities that would have refused treatment, we can estimate the number of control respondents with knowledge of government services who live in municipalities that would have accepted treatment, had it been offered: $250 - 30 = 220$. We also know the number of respondents with knowledge of government services who live in municipalities that accepted treatment, i.e., 370. The "effect of treatment on the treated" is then:

$$\frac{370}{250 - 30} = 1.7$$

Thus, the effect of treatment on the treated increased knowledge of government services by an estimated 70 percent.

In experiments with non-compliance, calculating the effect of treatment on the treated is straightforward, whenever there is "single crossover" – that is, there are subjects who take the treatment whether assigned to treatment or control *or* subjects who take the control whether assigned to treatment or control, but not both. If there is "double crossover," there are also useful estimators available, but the situation is more complicated (see Angrist et al. 1996 or Freedman 2006: 706 and passim).

Single-crossover seems like situation most likely to arise for many USAID programs. In this telecenter example, while some municipalities assigned to treatment may take the control (i.e., refuse to participate), municipalities assigned to control may have difficulty obtaining the treatment. Estimating the effect of treatment on the treated is therefore likely to be a useful tool for analyzing many experimental evaluations of USAID interventions.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444-455.

CEUDAP 2007a. "Measuring democracy: A survey and proposal." Draft document by John Gerring, to be included in CEUDAP report. May 7, 2007.

Collier, David, Henry E. Brady, and Jason Seawright. 2004. "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology." Chapter 13 in *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman & Littlefield.

Dabrowska, DM and TP Speed. 1990. "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles." English translation of Jerzy Neyman (1923), "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes." *Rocznici Nauk Rolniczych* 10: 1-51, in Polish. *Statistical Science* 5: 465-80 (with discussion).

Dunning, Thad. Forthcoming. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly*. Previous version presented at the annual meetings of the American Political Science Association, Washington D.C. Aug. 31-Sept. 3, 2005.

Eaton, Kent. 2007. "Backlash in Bolivia: Regional Autonomy as a Reaction against Indigenous Mobilization." *Politics and Society* 35 (1): 1-32.

Fearon, James. 1991. "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* 43: 169-95.

Finkel, Steve, Anibal Pérez Liñán and Mitchell A. Seligson. Forthcoming. "The Effects of U.S. Foreign Assistance on Democracy Building, 1990-2003." *World Politics*.

Freedman, David. 1991. "Statistical Models and Shoe Leather." In P.V. Marsden, ed., *Sociological Methodology,* Vol. 21. Washington, D.C.: The American Sociological Association.

Freedman, David A. 2006. "Statistical models for causation: What inferential leverage do they provide?" *Evaluation Review* 30: 691-713.

Freedman, D. A., D. B. Petitti, and J. M. Robins. 2004. "On the efficacy of screening for breast cancer." *International Journal of Epidemiology* 33: 43–73. Correspondence, pp. 1404–6.

GRADE. 2003. "Linea de Base Rápida: Gobiernos Subnacionales e Indicadores del

Desarrollo, Perfiles de siete regiones en la etapa inicial del proceso de decentralización." (GRADE Rapid Baseline). Preliminary report, Lima, Peru. May 2003.

LAPOP. 2007. "Cultural política de la democracia en el Perú: 2006." Barómetro de la Américas, Peru." Latin American Public Opinion Project (LAPOP) at Vanderbilt University and the Instituto de Estudios Peruanos. Mitch Seligson, Julio F. Carrión, and Patricia Zárate. Lima, Peru.

Neyman, Jerzy. 1923. "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes." *Rocznici Nauk Rolniczych* 10: 1-51, in Polish. English translation by DM Dabrowska and TP Speed (1990), *Statistical Science* 5: 465-80 (with discussion).

Posner, Daniel N. 2004. "The Political Salience of Cultural Difference: Why Chewas and Tumbukas Are Allies in Zambia and Adversaries in Malawi." *American Political Science Review* 98 (4): 529-545.

PRODES 2006. "Pro-Decentralization Performance Monitoring Plan (PMP)." Contract 527-C-00-03-00049-00, USAID/Peru and ARD, Inc. Original: 15 November 2003. Revised 22 November 2004, 15 April 2005, and 07 March 2006.

PRODES 2007. "Pro-Decentralization Performance Monitoring Plan (PMP)." 5th Year Option Period, 06 February 2007-05 February 2008. Contract 527-C-00-03-00049-00, USAID/Peru and ARD, Inc.

Rubin, Donald. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized studies." *Journal of Educational Psychology* 66: 688-701.

Snow, John. 1855. *On the Mode of Communication of Cholera.* Churchill, London. Reprinted in *Snow on Cholera*, London: Humphrey Milford: Oxford University Press, 1936.

SUNY. 2004. "Final Technical Report: Developing Skills of the Peruvian Congress." State University of New York (SUNY), Center for International Development. January 30, 2004.

USAID/Peru. 2002. "Request for Proposals (RFP) No. 527-P-02-019, Strengthening of The Decentralization Process and Selected Sub-National Governments in Peru ('the Pro-Decentralization Program')." Lima, Peru.

USAID. 1998. Handbook of Democracy and Governance Indicators. Technical Publication Series, Center for Democracy and Governance, USAID. August 1998.