

IMPROVING DEMOCRACY ASSISTANCE

**Building Knowledge Through
Evaluations and Research**



**NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES**

IMPROVING DEMOCRACY ASSISTANCE

**Building Knowledge Through
Evaluations and Research**

Committee on Evaluation of USAID Democracy Assistance Programs

Development, Security, and Cooperation
Policy and Global Affairs

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Contract No. DFD-C-00-06-00091-0 between the National Academy of Sciences and the U.S. Agency for International Development. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number-13: 978-0-309-11736-4

International Standard Book Number-10: 0-309-11736-4

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, <http://www.nap.edu>.

Copyright 2008 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

**COMMITTEE ON THE EVALUATION OF USAID
DEMOCRACY ASSISTANCE PROGRAMS**

Jack A. Goldstone (*Chair*), George Mason University, Fairfax, VA
Larry Garber, New Israel Fund, Washington, DC
John Gerring, Boston University, Boston, MA
Clark C. Gibson, University of California, San Diego, La Jolla, CA
Mitchell A. Seligson, Vanderbilt University, Nashville, TN
Jeremy Weinstein, Stanford University, Stanford, CA

Staff

Jo L. Husbands, Senior Project Director
Paul Stern, Board Director
Tabitha Benney, Senior Program Associate*
Rita Guenther, Senior Program Associate

*Until August 2007.

Preface

Since September 11, 2001, the Bush administration has made support for democracy one of the major pillars of U.S. security policy. The U.S. Agency for International Development (USAID) has been providing democracy assistance to countries around the world for over 25 years and has invested substantially in a variety of programs in diverse political situations. To better understand the impact of its democracy assistance efforts, the agency launched the Strategic and Operational Research Agenda (SORA). As part of SORA's work, USAID asked the National Research Council to prepare a report on how best to evaluate USAID democracy and governance (DG) programs.

The National Academies appointed an ad hoc committee to work on this report, including scholars with long experience and varied methodological approaches to the study of democracy and democratization, and a former USAID mission director with field experience in implementing DG programs. I extend my deepest personal thanks to each of them for their many intellectual contributions to the committee's work and for the time and effort they gave to the report. It was a pleasure to work with such outstanding colleagues.

To fulfill the mission given to the Academies, additional scholars were called on to help the committee examine key methodological issues in evaluating the impact of DG assistance. The committee's deliberations with these scholars included a conference in Boston, Massachusetts, on issues in measuring democracy and a conference in Stanford, California, on how case studies of democratization and democracy assistance could

inform DG programming. The committee also contracted with several expert consultants in the design and implementation of program evaluations, from both the academic and policy implementation spheres, to visit several USAID missions to examine the feasibility and scope for developing impact evaluations of DG projects in the field. The committee owes great thanks to these scholars for their contributions to this report. (A full listing of participants and consultants is given in Appendixes B through E.)

The committee spent many sessions discussing evaluation procedures with representatives of USAID and former and current contractors for the agency's DG programs. The committee is grateful to USAID for providing the time and assistance to set up these meetings and for their willingness to work through ideas and opportunities with the committee. In particular, the committee thanks three USAID officials who were the primary contacts throughout the project and whose support and advice were critical to the success of the committee's work: Margaret Sarles, chief of the Strategic Planning and Research Division, and two members of the SORA staff—David Black, who served as project officer, and Mark Billera. David and Mark also accompanied the committee's teams on their field visits. The goal was not merely to recommend abstract "ivory tower" ideas regarding project evaluations but to learn from USAID and provide recommendations that would be feasible in the field and useful on a variety of levels for USAID planning and program implementation.

The committee particularly wants to thank the USAID missions in Albania, Peru, and Uganda, who hosted committee members, staff, and consultants, and the USAID Washington officers who helped arrange those visits. The field visits were invaluable in determining how actual DG programs were being evaluated and learning how USAID staff and consultants could develop different evaluation designs for current and forthcoming USAID programs.

The committee also thanks the members of the National Research Council staff who provided substantive and administrative support for the project. Jo Husbands served as project director and helped guide the committee through dozens of meetings, lengthy deliberations, and the administrative hurdles of carrying out the committee's ambitious goals. She also made substantial contributions to the many drafts of the report. Paul Stern offered sage advice throughout the process, particularly on methodological and measurement issues. Rita Guenther and Tabitha Benney provided research and administrative support, along with immense energy and good cheer. Rita also took the lead in drafting the report summarizing the three field visits and drafted several sections of the report. Three of them also took part in the field visits—Jo in Albania and Uganda, Rita in Albania and Peru, and Tabitha in Peru.

At a time when democracy assistance is becoming ever more important as part of international and U.S. policies to assist developing nations, build peace, and reduce conflict, the committee hopes this report can serve as a practical guide for policymakers and USAID mission staff. Foreign assistance donors and aid organizations in a variety of areas are demanding better proof of results and more certain knowledge on which to build future assistance programs. The committee provides recommendations on how USAID can design its activities to gain greater knowledge of which DG projects are most effective in the field and how to use that knowledge—drawing on both internal experience and outside expertise—to guide and improve future democracy assistance. It is hoped that the recommendations in this report will lead to not only more effective programs to assist the emergence and stabilization of democracies but also the adoption of evaluation methods that will improve aid effectiveness throughout the domain of U.S. foreign assistance.

Jack A. Goldstone
Chair

Acknowledgments

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report: Kenneth Bollen, University of North Carolina, Chapel Hill; Valerie Bunce, Cornell University; Susan Hyde, Yale University; Robert Keohane, Princeton University; David Laitin, Stanford University; Carol Lancaster, Georgetown University; Ruth Levine, Center for Global Development; Michael Lund, Management Systems International; Gerald Munck, University of Southern California; Barbara Torrey, Population Reference Bureau, Inc.; and Nicholas van de Walle, Cornell University.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by Charles Tilly, Columbia University, and Enriqueta C. Bond, Burroughs Wellcome Fund. Appointed by the National Research Council, they were respon-

sible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

Contents

Summary	1
1 Democracy Assistance and USAID	17
U.S. Democracy Assistance: A Brief Introduction, 17	
Democratic Development and Democracy Assistance: What Do We Know?, 23	
USAID’s Request to the National Research Council, 27	
Report Overview, 32	
References, 38	
2 Evaluation in USAID DG Programs: Current Practices and Problems	43
Introduction, 43	
Current Evaluation Practices in Development Assistance: General Observations, 44	
Current Policy and Legal Framework for USAID DG Assessments and Evaluations, 53	
Three Key Problems with Current USAID Monitoring and Evaluation Practices, 58	
Conclusions, 66	
References, 67	
3 Measuring Democracy	71
Introduction, 71	

- Problems with Extant Indicators, 73
 A Disaggregated Approach to Measurement at the
 Country Level, 83
 Conclusions, 94
 References, 95
- 4 Learning from the Past: Using Case Studies of Democratic
 Transitions to Inform Democracy Assistance 99**
- Introduction, 99
 Case Study Designs and Methods, 100
 Insights from Current Research: Results of a Conference of
 Case Study Specialists on Democracy, 102
 A Multicase Study Design to Generate and Investigate Strategic
 Hypotheses Regarding Democracy Assistance, 112
 Conclusions, 116
 References, 117
- 5 Methodologies of Impact Evaluation 119**
- Introduction, 119
 Importance of Sound and Credible Impact Evaluations for
 DG Assistance, 120
 Plan of This Chapter, 124
 Points of Clarification, 124
 Internal Validity, External Validity, and Building
 Knowledge, 127
 A Typology of Impact Evaluation Designs, 132
 Examples of the Use of Randomized Evaluations in
 Impact Evaluations of Development Assistance
 (Including DG Projects), 144
 References, 148
- 6 Implementing Impact Evaluations in the Field 151**
- Introduction, 151
 Field Visits to USAID Missions, 152
 Employing Randomized Impact Evaluations for USAID
 DG Projects in the Field, 154
 Challenges in Applying Randomized Evaluation to DG
 Programs, 168
 Conclusions, 175
 References, 176

7	Additional Impact Evaluation Designs and Essential Tools for Better Project Evaluations	177
	Introduction, 177	
	How Often Are Randomized Evaluations Feasible?, 178	
	Designing Impact Evaluations When Randomization Is Not Possible, 181	
	What to Do When There Is Only One Unit of Analysis, 192	
	Conclusions, 196	
	References, 197	
8	Creating the Conditions for Conducting High-Quality Evaluations of Democracy Assistance Programs and Enhancing Organizational Learning	199
	Introduction, 199	
	Issues in Obtaining High-Quality Impact Evaluations, 199	
	Improving Organizational Learning, 208	
	Conclusions, 216	
	References, 217	
9	An Evaluation Initiative to Support Learning the Impact of USAID's DG Programs	219
	Introduction, 219	
	Providing Leadership and Strategic Vision, 220	
	Implementing the Vision: The Evaluation Initiative, 222	
	Agenda for USAID and SORA, 228	
	Role of Congress and The Executive Branch, 230	
	Conclusions, 232	
	References, 232	
	Glossary	235
	Appendixes	
A	Biographical Sketches of Committee Members	243
B	Committee Meetings and Participants	247
C	Measuring Democracy	259
D	Understanding Democratic Transitions and Consolidation from Case Studies: Lessons for Democracy Assistance	285
E	Field Visit Summary Report	289
F	Voices from the Field: Model Questionnaire	315

Summary

BACKGROUND

Over the past 25 years, the United States has made support for the spread of democracy to other nations an increasingly important element of its national security policy. Many other multilateral agencies, countries, and nongovernmental organizations (NGOs) also are involved in providing democracy assistance. These efforts have created a growing demand to find the most effective means to assist in building and strengthening democratic governance under varied conditions.

Within the U.S. government the U.S. Agency for International Development (USAID) has principal responsibility for providing democracy assistance. Since 1990, USAID has supported democracy and governance (DG) programs in approximately 120 countries and territories, spending an estimated total of \$8.47 billion (in constant 2000 U.S. dollars) between 1990 and 2005. The request for DG programs for fiscal year 2008 was \$1.45 billion, which includes some small programs in the U.S. Department of State.

Despite these substantial expenditures, our understanding of the actual impacts of USAID DG assistance on progress toward democracy remains limited—and is the subject of much current debate in the policy and scholarly communities. Admittedly, the realities of democracy programming are complicated, given the emphasis on timely responses in politically sensitive environments and flexibility in implementation to account for fluid political circumstances. These realities pose particular challenges for the evaluation of democracy assistance programs. Nonethe-

less, USAID seeks to find ways to determine which programs, in which countries, are having the greatest impact in supporting democratic institutions and behaviors and how those effects unfold. To do otherwise would risk making poor use of scarce funds and to remain uncertain about the effectiveness of an important national policy.

Yet USAID's current evaluation practices do not provide compelling evidence of the impacts of DG programs. While gathering valuable information for project tracking and management, these evaluations usually do not collect data that are critical to making the most accurate and credible determination of project impacts—such as obtaining baseline measures of targeted outcomes before a project is begun or tracking changes in appropriately selected (or assigned) comparison groups to serve as a control or reference group.

USAID has been seeking better evidence for the effects of its DG projects. In 2000 the Office of Democracy and Governance created the Strategic and Operational Research Agenda (SORA). Under SORA, USAID has commissioned studies of its DG evaluations and underwritten a recent cross-national study of the effects of its democracy assistance programs since 1990. A very encouraging finding from that study is that democracy assistance *does* matter for democratic progress. The study (Finkel et al 2007; see also the second-phase study, Finkel et al 2008) found that, when controlling for a wide variety of other factors, higher levels of democracy assistance are, on average, associated with movement to higher levels of democracy. These results provide the clearest evidence to date that democracy assistance contributes toward achieving its desired goals.

Unfortunately, it is also true that in a number of highly important cases—such as Egypt and post-Soviet Russia—large volumes of democracy assistance have yielded disappointing results. In addition to knowledge about general effects, USAID needs to know the positive or negative effects of specific projects and why DG assistance has been more successful in some contexts than in others. SORA turned to the National Research Council (NRC) for assistance in how to gain greater insight into which democracy assistance projects are having the greatest impacts. This report is intended to provide a road map to enable USAID and its partners to build, absorb, and act on improved knowledge about assisting the development of democracy in a variety of contexts.

CHARGE TO THE COMMITTEE

The USAID Office of Democracy and Governance asked the NRC for help in developing improved methods for learning about the effectiveness and impact of its work, both retrospectively and in the future. Specifically, the project is to provide:

1. A refined and clear overall research and analytic design that integrates the various research projects under SORA into a coherent whole in order to produce valid and useful findings and recommendations for democracy program improvements.

2. An operational definition of democracy and governance that disaggregates the concept into clearly defined and measurable components.

3. Recommended methodologies to carry out retrospective analysis. The recommendations will include a plan for cross-national case study research to determine program effectiveness and inform strategic planning. USAID will be able to use this plan as the basis of a scope of work to carry out comparative retrospective analysis, allowing the agency to learn from its 25 years of investment in DG programs.

4. Recommended methodologies to carry out program evaluations in the future. The recommendations for future analysis will focus on more rigorous approaches to evaluation than currently used to assess the impact of democracy assistance programming. They should be applicable across the range of DG programs and allow for comparative analysis.

5. An assessment of the feasibility of the final recommended methodologies within the current structure of USAID operations and defining policy, organizational, and operational changes in those operations that might improve the chances for successful implementation.

OVERALL RESEARCH AND ANALYTIC DESIGN

In response to the first charge, the committee unanimously recommends a four-part strategy for gaining increased knowledge to support USAID's DG policy planning and programming. These are:

Recommendation 1: *Undertaking a pilot program of impact evaluations designed to demonstrate whether such evaluations can help USAID determine the effects of its DG projects on targeted policy-relevant outcomes.* A portion of these impact evaluations should use randomized designs since, where applicable and feasible, they are the designs most likely to lead to reliable and valid results in determining project effects and because their use in DG projects has been limited. USAID should begin the pilot program by focusing on a few widely used DG program categories. The pilot evaluations should not supplant current evaluations and assessments, but impact evaluations could gradually become a more important part of USAID's portfolio of monitoring and evaluation (M&E) activities as the agency gains experience with such evaluations and determines their value. (See Chapters 5 through 7 for a discussion of impact evaluations and how they might be applied to DG projects and Chapter 9 for the committee's recommendations.)

Recommendation 2: *Developing more transparent, objective, and widely accepted indicators of changes in democratic behavior and institutions at the sectoral level*—that is, at the level of such sectors as the rule of law, civil society, government accountability, effective local government, and quality of elections. Current aggregate national indicators of democracy, such as Freedom House or Polity scores, are neither at the right level for identifying the impacts of particular USAID DG projects nor accurate and consistent enough to track modest or short-term movements of countries toward or away from greater levels of democracy. (See Chapter 3.)

Recommendation 3: *Using more diverse and theoretically structured clusters of case studies of democratization and democracy assistance to develop hypotheses to guide democracy assistance planning in a diverse range of settings.* Whether USAID chooses to support such studies or gather them from ongoing academic research, it is important to look at how democracy assistance functions in a range of different initial conditions and trajectories of political change. Such case studies should seek to map out long-term trajectories of political change and to place democracy assistance in the context of national and international factors affecting those trajectories, rather than focus mainly on specific democracy assistance programs. (See Chapter 4.)

Recommendation 4: *Rebuilding USAID’s institutional mechanisms for absorbing and disseminating the results of its work and evaluations, as well as its own research and the research of others, on processes of democratization and democracy assistance.* In recent years, USAID has lost much of its capacity to assess the impact and effectiveness of its programs. Without an active program of organizational learning so that senior personnel and DG officers have structured opportunities to discuss the results of pilot evaluations, compare their experiences with DG programs, and discuss the research carried out by USAID and especially other scholars, implementers, and donors, the fruits of the committee’s first three recommendations will not be usefully integrated with the experience of DG officers in a way that will improve DG program planning, design, and outcomes. (See Chapters 8 and 9.)

DISCUSSION AND STRATEGIES FOR IMPLEMENTATION

The following sections provide more detail on the reasons behind these recommendations and discuss organizational issues at USAID that will affect the agency’s ability to implement them.

Recommendation 1: Undertaking a Pilot Program of Impact Evaluations

Charges 4 and 5 asked the committee to recommend methodologies for future program evaluations and to evaluate their feasibility. These issues are addressed first, however, because the committee believes that, among the charges it was given, improving USAID's ability to more precisely ascertain the effects of future DG programs has more potential to build knowledge of what works best in DG programming than either retrospective analyses (given the limits found in the collection of data on past DG projects) or improving the definition of democracy. The committee thus investigated USAID's current evaluation methods and explored a range of designs for improved evaluations that could be applied to DG projects. The committee also commissioned teams of consultants to visit three diverse missions—in Albania, Peru, and Uganda—to assess the feasibility of applying those designs—in particular impact evaluations—to actual ongoing or planned DG projects. Of course, these evaluations, like all of USAID's evaluations and research, must be part of a broader learning strategy if the agency is to benefit; these organizational aspects are discussed separately below.

What Are Impact Evaluations?

Most current evaluations of USAID DG projects, while informative and serving varied purposes for project managers, lack the designs or data needed to provide compelling evidence of whether those projects had their intended effects. An *impact evaluation* aims to separate the effects of a specific DG project from the vast range of other factors affecting the progress of democracy in a given country and thus to make the most precise and credible determination of how much DG projects contribute to desired outcomes.

As the committee uses the term, what distinguishes an impact evaluation is the effort to determine what would have happened in the absence of the project by using comparison or control groups, or random assignment of assistance across groups or individuals, to provide a reference against which to assess the observed outcomes for groups or individuals who received assistance. Randomized designs offer the most accuracy and credibility in determining program impacts and therefore should be the first choice, where feasible, for impact evaluation designs. However, such designs are not always feasible or appropriate, and a number of other designs also provide useful information to determine the impact of many different kinds of assistance projects. For example, when there is only one group or institution receiving assistance, comparisons may be made across time by using a set of carefully timed measures before and

after the project while controlling statistically for long-term trends or key events. Impact evaluations are designed according to standard protocols of evaluation research; yet the choice of a particular design and decisions about how to adapt the design to a particular project require skilled craftsmanship as much as science.

Current Approaches to Evaluation in USAID

The committee's review of current approaches to the evaluation of development assistance in general, and USAID DG programs in particular, found that:

- Very few of the evaluations undertaken by international or multilateral development and democracy donors are designed as impact evaluations. There are signs that this is changing as some donors and international agencies are beginning to implement new approaches to evaluation. The Millennium Challenge Corporation and the World Bank in particular have undertaken efforts to increase the use of randomized designs in evaluations of their economic assistance and anticorruption projects. A few NGOs also have undertaken randomized impact evaluations of their democracy assistance efforts.

- Within USAID the number of evaluations has declined for *all* types of assistance programs. The evaluations undertaken for DG programs generally focus on implementation and management concerns and have not collected the data needed for sound impact evaluations. For example, most past evaluations of DG projects have not made comparable baseline and postproject data measurements on key outcomes, and almost all past evaluations lacked data on comparison groups that did not receive assistance. This makes it nearly impossible to develop a retrospective analysis from the data in those evaluations to accurately determine the effects of DG programs.

- There is a tendency, at one and the same time, to evaluate democracy projects mainly in terms of very proximate outcome measures that mainly assess how well the project was implemented and yet to judge the ultimate success of DG projects by whether they coincide with changes in country-level measures of national democracy such as Freedom House scores. Neither course best serves USAID's interests in determining the effects of its DG programs. Those effects are best judged by focusing on policy-relevant objectives at the local or sectoral level that are plausible outcomes of those projects.

- Once research and evaluation are completed, there are few organizational mechanisms for broad discussion among DG officers or for

integration of research and evaluation findings with the large range of analysis being carried on outside the agency.

- DG officials are genuinely interested in procedures that will help them better learn and demonstrate the impact of their projects. Yet there is considerable concern among many at USAID regarding whether missions would gain from designing or implementing rigorous impact evaluations, especially those using randomized assignments. This is mostly due to deep skepticism as to the applicability of this methodology to DG programs but also to the overall decline in support for evaluations within USAID, to a lack of specific expertise on impact evaluation design, and to issues in contracting timetables and procedures that discourage adoption of what is perceived as a more complicated approach to evaluation.

- More generally, while there are many calls from policymakers, USAID officials, and other international and national agencies and donors to better determine the effects of DG programs, there is also widespread skepticism regarding whether impact evaluations will, in fact, provide that information. One member of the committee, Larry Garber, emphatically shares these concerns. Among both scholars and policy professionals, skeptics worry that the designs for impact evaluations will prove too cumbersome or inflexible to work in fluid and politically sensitive conditions in the field; that such evaluations will be too costly or time-consuming; or that such studies, in particular randomized designs, are either unethical for or ill suited to the actual projects being carried out in DG programs.

Feasibility of Impact Evaluations for DG Projects

Recognizing the need to take such concerns seriously, the committee examined a wide range of impact evaluation designs and worked with DG officers at several missions to assess the feasibility of such designs for their current or planned activities. The committee's field studies found that a much larger portion of USAID's DG programs than expected—forming roughly half of the projects that were examined in Uganda and several projects in Peru and Albania—appear to be amenable, in the view of the committee's consultants, to randomized assignment designs. Nor did these designs necessarily require major departures from current program procedures. Often just more attention to how programs were rolled out or allocated among groups scheduled to receive assistance, combined with measurements on both the groups currently receiving assistance and those scheduled to receive it in the future, would create a reasonable randomized assignment design. In cases where randomized assignment designs were not feasible, the field teams were able to develop other

designs that could offer a significant improvement in the ability to assess project effectiveness.

In addition, the committee found that many of the surveys that USAID is already carrying out provide excellent baseline and comparison data for DG projects; thus the data for impact evaluations that use matched or adjusted comparison groups (rather than randomization) are in some cases already being collected and could be utilized for little additional cost.

The field teams thus concluded that it was quite feasible, at least in theory, to conduct high-quality impact evaluations of varied designs that will help USAID better discern the impacts of its DG programs. However, the committee knows that there is much skepticism regarding these procedures and, in particular, concerns—noted by Mr. Garber and by others in the democracy assistance donor community—about whether the complexity and sensitivity of DG programs will permit sound impact evaluations, especially those using randomized assignments, to be carried out. Therefore the full committee agreed that the value of such impact evaluations will have to be demonstrated in USAID's own experience.

Strategies for Implementation

- **The committee unanimously recommends that USAID move cautiously but deliberately to implement pilot impact evaluations of several carefully selected projects, including a portion with randomized designs, and expand the use of such impact evaluations as warranted by the results of those pilot evaluations and the needs expressed by USAID mission directors.**

- **Moreover, the committee recommends that these pilot evaluations be undertaken as part of a DG evaluation initiative with senior leadership that will also focus on improving USAID's capacity to undertake impact evaluations and make resources and expertise available to mission directors seeking to learn about and apply impact evaluations to their projects. This DG evaluation initiative is described in more detail below.**

Recommendation 2: Developing Better Sectoral-Level Indicators Measuring Democracy

In response to Charge 2, the committee reviewed the most widely used indicators of a country's overall democratic status and considered a number of alternative approaches to developing an operational definition of democracy. This led to four key findings:

- The concept of democracy cannot, in the present state of scientific knowledge of democracies and democratization, be defined in an authoritative (nonarbitrary) and operational fashion. It is an inherently multidimensional concept, and there is little consensus over its attributes. Definitions range from minimal—a country must choose its leaders through contested elections—to maximal—a country must have universal suffrage, accountable and limited government, sound and fair justice and extensive protection of human rights and political liberties, and economic and social policies that meet popular needs. Moreover, the definition of democracy is itself a moving target; definitions that would have seemed reasonable at one time (such as describing the United States as a democracy in 1900 despite no suffrage for women and major discrimination and little office-holding among minorities) are no longer considered reasonable today.

- Existing empirical indicators of overall democracy in a country suffer from flaws that include problems of definition and aggregation, imprecision, measurement errors, poor data coverage, and a lack of agreement among scales intended to measure the same qualities. There is thus no way to utilize existing macro-level indicators in a way that provides sound policy guidance or reliably tracks modest or short-term changes in a country's democratic status. Existing indicators work best simply to roughly categorize countries as "fully democratic," "authoritarian," or "mixed or in between" and to identify large-scale or long-term movements in levels of democracy. They are particularly weak in assessing differences among the nondemocratic and mixed regimes that are the most important settings for USAID's DG work.

- By contrast, indicators focused on specific sectors of democracy in a country (the sectoral level) would help USAID (1) track trends across various dimensions of democracy through time, (2) make precise comparisons across countries and regions, (3) understand the components and possible sequences of democratic transition, (4) analyze causal relationships (e.g., between particular facets of democracy and economic growth), and (5) assess the democratic profile (i.e., strengths and weaknesses across various dimensions of democracy) of countries where USAID operates.

- While the United States, other donor governments, and international agencies that are making policy in the areas of health or economic assistance are able to draw on databases that are compiled and updated at substantial cost by government or multilateral agencies mandated to collect such data, no comparable source of data on democracy at either the macro or sectoral level currently exists. Data on democracy are instead currently compiled by various individual academics on irregular and shoestring budgets, or by NGOs or commercial publishers, using different definitions and indicators of democracy.

Strategies for Implementation

These findings have led the committee to make a recommendation that committee members believe would significantly improve USAID's (and others') ability to track countries' progress and make the type of strategic assessments that will be most helpful for DG programming.

- **USAID and other policymakers should explore making a substantial investment in the systematic collection of democracy indicators at a disaggregated sectoral level—focused on the components of democracy rather than (or in addition to) the overall concept.** If they wish to have access to data on democracy and democratization comparable to the data relied on by policymakers and foreign assistance agencies in the areas of public health or trade and finance, a substantial government or multilateral effort to improve, develop, and maintain international data on levels and detailed aspects of democracy would be needed. This should not only involve multiple agencies and actors in efforts to initially develop a widely accepted set of sectoral data on democracy and democratic development but should also seek to institutionalize the collection and updating of democracy data for a broad clientele, along the lines of the economic, demographic, and trade data collected by the World Bank, the United Nations, and the International Monetary Fund.

- Although creating better measures at the sectoral level to track democratic change is a long-term process, there is no need to wait on such measures for determining the impact of USAID's DG projects. USAID has already compiled an extensive collection of policy-relevant indicators to track specific changes in government institutions or citizen behavior, such as levels of corruption, levels of participation in local and national decision making, quality of elections, professional level of judges or legislators, or the accountability of the chief executive. **Since these are, in fact, the policy-relevant outcomes that are most plausibly affected by DG projects, the committee recommends that measurement of these factors rather than sectoral-level changes be used to determine whether the projects are having a significant impact on the various elements that compose democratic governance.**

Recommendation 3: Using Case Studies of Democratization and Democracy Assistance

The third charge to the committee was to recommend a plan for comparative historical case studies of DG assistance. A clustered set of case studies, tracing the processes through which advances toward democracy were made from various sets of initial conditions, is an appropriate mode of investigation for these issues. Such case studies could be particularly

valuable in mapping out varied trajectories of political development and identifying the role that democracy assistance could play in such trajectories in relation to various actors and events.

Nonetheless, committee members were unable to agree on a firm recommendation that USAID should invest its own funds in such case studies since substantial case study research on democratization is being undertaken by academics and NGOs. To learn more about the role of its DG assistance projects in varied conditions and their role in varied trajectories of democratization, USAID could seek to gain from ongoing academic research. Since much potentially relevant academic research is not written for a policy audience, however, USAID would need to structure its interactions with researchers to ensure that it gains useful and relevant information.

Strategies for Implementation

- **If USAID decides to invest in supporting case study research, the committee recommends using a competitive proposal solicitation process to elicit the best designs. USAID should not specify a precise case study design, but instead should specify key criteria that proposals must meet.** These should include (1) the criteria for choosing cases should be explicit and theoretically driven; (2) the cases should include a variety of initial conditions or contexts in which USAID DG projects operate; (3) the cases should include at least one, if not several, countries in which USAID and other donors have made little or no investment in DG projects; and (4) the cases should include countries with varied outcomes regarding democratic progress or stabilization.

- In addition to case studies, a variety of other research methods, both formal and informal (including debriefings of USAID field officers, statistical analyses of international data, and surveys) can shed light on patterns of democratization as well as how DG projects actually operate in the field and how they are received. **USAID should include these varied sources of information as part of the regular organizational learning activities the committee recommends next.**

Recommendation 4: Rebuilding USAID's Institutional Mechanisms for Learning

Regardless of whether USAID conducts many or fewer impact evaluations and contracts for case studies or works with case studies funded by think tanks or other organizations, little of what is learned will effectively guide and improve DG programming without some mechanism within USAID for learning from its own and others' research on democracy and democratization. For USAID to benefit from this committee's proposed

pilot study of impact evaluations, it will need to have regular means of disseminating the results of those and other evaluations throughout the agency and discussing the lessons learned from them. For USAID to benefit from ongoing academic research and the studies of DG assistance being undertaken by think tanks and NGOs, it will be necessary for the agency to organize regular structured interactions between such researchers and its DG staff.

While it will take some time for USAID to learn from undertaking the pilot impact evaluations, it will gain immediately from augmenting its overall learning activities and increasing opportunities for DG staff to actively engage with current research and studies on democratization. Though some committee members believe that the impact evaluations will be more novel and instructive than most current case study and policy reports on democratization, several committee members wish to emphasize the considerable value to policymakers and DG officers of the many books, articles, and reports that have been prepared in recent years by academics, think tanks, and practitioners. Whatever the methodological flaws of these case studies and process evaluations from a rigorous social sciences perspective, the committee notes that this expanding literature has provided important lessons and insights for crafting effective DG programs. Thus the committee is unanimous in finding that a renewed emphasis on engaging USAID DG personnel in discussion and analysis of current research on democratization and democracy assistance—including both varied types of evaluations and a broad range of scholarship—would be worthwhile and should begin even before the pilot evaluations have been completed.

Unfortunately, in recent years USAID has substantially reduced its institutional mechanisms for creating, disseminating, and absorbing knowledge. The Center for Development Information and Evaluation (CDIE), which served as the hub of systemic evaluation for USAID aid projects, has been dissolved. Moreover, USAID's support of conferences and learning activities for mission directors and DG staff to share experiences and discuss the latest research has declined. And although central collection of evaluations is already a requirement, in practice much useful information, including evaluations and other project documents, survey data and reports, and mission director and DG staff reports, remains dispersed and difficult to access.

Strategies for Implementation

Rebuilding organizational learning capacity within USAID will require a number of steps, some minor and some potentially involving major shifts in organizational procedures. The committee thus recom-

mends that the steps below be undertaken by a special DG evaluation initiative led by a senior policymaker or official within USAID who will have the ability to recommend agency-wide changes, as many of the obstacles to improved learning about DG programs stem from agency-wide procedures and organizational characteristics. While in some ways this will replace the capabilities lost with CDIE, in some ways the committee hopes the new initiative will go beyond that.

The committee's charge is limited to recommendations for improving USAID's ability to evaluate its DG projects, but the committee notes that there could be advantages to making this an agency-wide initiative. USAID implements social programs in many parts of the agency, so the changes the committee recommends could yield much wider benefits.

A DG EVALUATION INITIATIVE

In support of Recommendations 1 and 4, the committee recommends that USAID develop a five-year DG evaluation initiative, led by a senior USAID official and with special funding, for the following:

1. Undertaking Pilot Impact Evaluations

The committee strongly recommends that to accelerate the building of a solid core of knowledge regarding project effectiveness, the DG evaluation initiative should immediately develop and undertake a number of well-designed impact evaluations that test the efficacy of key project models or core development hypotheses that guide USAID DG assistance. A portion of these evaluations should use randomized designs, as these are the most accurate and credible means of ascertaining program impact. As randomized designs have also been the most controversial, especially in the DG area, it would be most valuable for the evaluation initiative to help USAID gain experience with and determine the value of these designs for learning the impacts of DG projects.

By key models the committee refers to programs that (1) are implemented in a similar form across multiple countries and (2) receive substantial funding (e.g., local government support, civil society, judicial training). By core hypotheses the committee refers to assumptions guiding USAID program design that, whether drawn from experience or prevailing ideas about how democracy is developed and sustained, have not been tested as empirical propositions.

2. Increasing USAID's Capabilities in Project Evaluation

Supporting the DG evaluation initiative with special, dedicated resources outside the usual project structure would be another signal of a strong commitment to change. It is also important that these

resources and accompanying expertise in evaluation design be made available to missions implementing DG programs, so that more rigorous evaluations become an opportunity for missions to gain support, rather than an additional unfunded burden. Any changes to M&E of DG programs will be carried out in the field by over 80 missions and hundreds of implementing partners. Even with the centralization of program and budget decision making undertaken in the foreign assistance reforms of 2006, USAID is a highly decentralized agency and mission staff have substantial discretion in how they implement and manage their programs. **The initiative should thus make its resources and expertise available to mission directors who want its support in conducting impact evaluations or otherwise changing their mix of M&E activities, in order to make the initiative an asset to the DG directors in the field rather than an additional unfunded burden.**

3. Providing Technical Expertise

In recent years as USAID has reduced the number of evaluations it conducts, the agency has also failed to hire experts in the latest evaluation practices to guide and oversee its contracting and research. **The committee recommends that USAID acquire sufficient internal expertise in this area to both guide an initiative at USAID headquarters and provide advice and support to field missions, as a key element of the initiative.**

4. Improving the Ease of Undertaking Impact Evaluations of DG Projects

While many evaluations are currently only sought well after a project has begun or even only after its completion, impact evaluations generally require before-and-after measures and data from comparison or control groups that should be designed into the program from its inception and often cannot be obtained at all once a program is well under way. Pressures to get projects under way, as well as many current contracting practices, thus work against implementing and sustaining impact evaluation designs. **One task of the DG evaluation initiative should be to address these issues and explore how to ease the task of undertaking impact evaluations within USAID's contracting and program procedures. The initiative should also examine incentives for both DG officers and project implementers to carry out sound impact evaluations of selected DG projects.**

5. Consider Creating a Social Sciences Advisory Group

To assist in the evaluation effort, the committee recommends that the administrator consider establishing a social sciences advisory group

for USAID. This group could play a useful role in advising on the design of the DG evaluation initiative, helping to work through issues that arise during implementation, and developing a peer review process for assessing the evaluations undertaken during the initiative.

6. Rebuilding Institutional Learning Capacity

This initiative should be guided by a policy statement outlining the strategic role of developing USAID as a learning organization in the democracy sector. The committee believes that increasing USAID's capacity to learn what works and what does not should include provisions for regular face-to-face interactions among DG officers, implementers, and outside experts to discuss recent findings, both from the agency's own evaluations of all kinds and studies by other donors, think tanks, and academics. Videoconferencing and other advanced technologies can be an important supplement, but personal contact and discussion would be extremely important to sharing experiences of success and failure as the evaluation initiative went forward. This includes lessons about both the effectiveness of DG projects and successes and failures in implementing impact evaluations.

Such meetings are especially important for ensuring that the varied insights derived from impact and process evaluations, academic studies, and examinations of democracy assistance undertaken by independent researchers, NGOs, think tanks, and other donors are absorbed, discussed, and drawn into USAID DG planning and implementation. While only USAID has the ability to develop and carry out rigorous evaluations of its projects' impacts, many organizations are carrying out studies of various aspects of democracy assistance, and USAID's staff can benefit from the wide range of insights, hypotheses, and "lessons learned" that are being generated by the broader community involved with democracy promotion.

Results of the Initiative

At the end of this five-year period, USAID would have:

- Practical experience in implementing impact evaluation designs that will indicate where such approaches are feasible, what the major obstacles are to wider implementation, and whether and how these obstacles can be overcome.
- Where the evaluations prove feasible, a solid empirical foundation for assessing the validity of some of the key assumptions that underlie DG projects and rigorous determinations of the impact of commonly used DG projects in achieving program goals.

- A core of expertise within USAID on the latest evaluation methods and practices.
- Institutionalized learning practices across the organization to keep officials engaged, informed, and up-to-date on the latest findings from within and outside USAID regarding democracy and democracy assistance.

CONCLUSION

The committee stresses that the goal of USAID should not be merely incremental improvement of its project evaluations, or funding additional case studies, but building the entire capacity of the agency to generate, absorb, and disseminate knowledge regarding democracy assistance and its effects. This will necessarily involve (1) gaining experience with varied impact evaluation designs, including randomized studies, to ascertain how useful they could be for determining the effects of DG projects; (2) focusing on disaggregated, sectoral-level measures to track democratic change; (3) expanding the diversity of case studies that are used to inform thinking on DG planning; and (4) adopting mechanisms and activities to support the active engagement of DG staff and mission personnel with new research on democratization and DG assistance.

REFERENCES

- Finkel, S.E., Pérez-Liñán, A., and Seligson, M.A. 2007. The Effects of U.S. Foreign Assistance on Democracy Building, 1990-2003. *World Politics* 59(3):404-439.
- Finkel, S.E., Pérez-Liñán, A., Seligson, M.A., and Tate, C.N. 2008. Deepening Our Understanding of the Effects of U.S. Foreign Assistance on Democracy Building: Final Report. Available at: <http://www.LapopSurveys.org>.

Democracy Assistance and USAID

U.S. DEMOCRACY ASSISTANCE: A BRIEF INTRODUCTION

The United States has been supporting democracy abroad for many decades. From Woodrow Wilson's efforts following World War I to the reconstruction of Germany and Japan after World War II, U.S. policymakers have aimed to create a world of democratic nations. During the Cold War and the current war on terrorism, efforts to foster democracy have been inconsistent or have clashed with other strategic goals, but the U.S. commitment to the growth of democracy abroad has been repeatedly expressed. Over the past 25 years, the United States has made assistance for the development of democracy in other nations a key element of its national security policy (see Box 1-1).

In recent years democracy assistance has become not merely a goal for diplomacy (although it remains that) but an increasingly frequent practical problem. A host of international and multilateral donor agencies and even military forces (both NATO and U.S.) have taken on the task of helping build democracies in highly challenging environments, including authoritarian and semiauthoritarian states, recently emerging and transitional democracies, and societies scarcely out of, or even in the midst of, violent conflicts (e.g., Ukraine, Bosnia, Egypt, Afghanistan, Iraq, Haiti, Democratic Republic of the Congo). U.S. efforts to assist the spread of democracy encompass a host of activities: diplomatic pressures, trade sanctions, economic development aid, military and political support for democratic forces, or in some cases (e.g., Zaire, Philippines) withdrawal of support for dictators.

BOX 1-1
Examples of U.S. Commitments to
Democracy Promotion, 1982-2006

"The objective I propose is quite simple to state: to foster the infrastructure of democracy, the system of a free press, unions, political parties, universities, which allows a people to . . . reconcile their own differences through peaceful means. . . . It is time that we committed ourselves as a nation—in both the public and private sectors—to assisting democratic development."

—**President Ronald Reagan**, "Speech at Westminster," June 8, 1982. Available at: <http://www.teachingamericanhistory.org/library/index.asp?documentprint=926>.

"Our interests are best served in a world in which democracy and its ideals are widespread and secure. We seek to . . . promote the growth of free, democratic political institutions as the surest guarantors of both human rights and economic and social progress."

—**National Security Strategy of the United States**, August 1991. Available at: <http://www.fas.org/man/docs/918015-nss.htm>.

"The best way to advance America's interests worldwide is to enlarge the community of democracies and free markets throughout the world."

—**President William J. Clinton**, "Statement on the National Security Strategy Report," July 21, 1994. Available at: <http://www.presidency.ucsb.edu/ws/index.php?pid=50525>.

"We will . . . use our foreign aid to promote freedom, . . . ensuring that nations moving toward democracy are rewarded for the steps they take [and] make freedom and the development of democratic institutions key themes in our bilateral relations."

—**National Security Strategy of the United States**, September 2002. Available at: <http://www.whitehouse.gov/nsc/nss.html>.

"I would define the objective of transformational diplomacy this way: To work with our many partners around the world to build and sustain democratic, well-governed states that will respond to the needs of their people—and conduct themselves responsibly in the international system."

—**Secretary of State Condoleezza Rice**, January 18, 2006. Available at: <http://www.state.gov/r/pa/prs/ps/2006/59339.htm>.

Role of the U.S. Agency for International Development

The day-to-day tasks of working with groups and individuals on the ground to help build democratic institutions and offer training and support to citizens, officials, and civil society organizations are assigned primarily to the U.S. Agency for International Development (USAID). USAID was created by executive order in 1961, following passage of the

Foreign Assistance Act, but its roots reach back to efforts such as the Marshall Plan to reconstruct Europe after World War II and the Food for Peace Program. Originally created to promote economic development, over the years the agency's mandate has expanded to include health, the environment, humanitarian assistance, conflict management and mitigation, and the promotion of democracy and good governance, as each of these has been deemed crucial to the overall U.S. foreign policy goals of improving the social and economic welfare of developing countries and increasing international peace and stability.

USAID's current democracy and governance (DG) activities date from the mid-1980s when a series of countries in Latin America, Asia, Africa, and then Central Europe and the former Soviet Union began the transition from various forms of authoritarian rule. Presidents Reagan, George H.W. Bush, and Clinton gave USAID the tasks of providing assistance to countries trying to develop democratic forms of government and creating programs to encourage other countries to embark on similar reforms. The administration of George W. Bush has continued and in some cases expanded this aid as a key element in its policy of "transformational diplomacy."

Behind efforts to support the spread of democracy promotion lies the belief that increasing democracy in developing nations will promote economic growth, diminish the risks of terrorism, and reduce the frequency of internal and international conflicts. Whether or not democracy actually has all of these effects, and under what conditions, is far from certain. As discussed further below, there is a substantial academic and policy debate on the merits of promoting democratic transitions (Goldstone and Ulfelder 2004, Halperin et al 2004, Mansfield and Snyder 2005, Ackerman 2006, Sanders and Halperin 2006, Epstein et al 2007). However, at present the international community, led mainly by democratic nations, continues to believe that helping nations transition to democracy is a significant route to promoting peace and economic development. This debate is far beyond the scope of this report, which will accept the goal of supporting democracy as a current aspect of policy and focus on how USAID can better assess whether its current efforts are having an impact on achieving that goal.

Since 1990, USAID has supported democracy programs in approximately 120 countries and territories with budgets ranging from tens of thousands to hundreds of millions of dollars. The most comprehensive analysis of USAID DG spending estimates total expenditures between 1990 and 2005 at \$8.47 billion in constant 2000 U.S. dollars (Azpuru et al. 2008). Total annual USAID DG expenditures currently run over \$1 billion; for fiscal year (FY) 2008 the request for DG, including both USAID and some much smaller amounts for the State Department, was \$1.45 billion,

with \$374 million allocated to Iraq and Afghanistan (Congressional Budget Justification [CBJ] 2008).¹

The programs are supported by hundreds of DG officers and other personnel in Washington and at overseas missions. As of 2004, DG comprised the agency's largest category of technical expertise among direct hire personnel at just over 400 (USAID 2006), although not everyone in this category is doing DG work at any given time.

Yet the funding of DG efforts, given their high priority for U.S. foreign policy and frequent mandate to help transform political systems into democracies, is relatively modest. In many countries, projects that are not strictly DG but that respond to related national needs may find a home under the DG umbrella, so the amount of effort actually focused on democracy building is smaller than may at first appear.² Moreover, DG funds comprise only a small portion of what the United States spends on its international engagements. The total FY2008 budget request for foreign assistance, which includes DG programs, was \$20.3 billion (CBJ 2008:1). Secretary of Defense Robert Gates (2007) argued in a speech at Kansas State University:

Funding for non-military foreign-affairs programs has increased since 2001, but it remains disproportionately small relative to what we spend on the military and to the importance of such capabilities. Consider that this year's budget for the Department of Defense—not counting operations in Iraq and Afghanistan—is nearly half a trillion dollars. The total foreign affairs budget request for the State Department is \$36 billion, less than what the Pentagon spends on health care alone.

This means that direct funding for democracy assistance by the United States constitutes less than 10 percent of U.S. spending on foreign assistance (most of which is for economic and humanitarian aid), about 4 percent of total nonmilitary spending on foreign affairs, and less than one-quarter of 1 percent of what is spent by the U.S. military. Put another way, the entire U.S. DG budget request for \$1.45 billion for FY2008 for worldwide efforts to transform countries into stable democracies is about one-tenth the annual budget request of the State of California's Depart-

¹One result of the consolidated budgeting process instituted as part of the foreign assistance reforms described in the next chapter is that, at least for the FY2008 request, it is not possible to break USAID out from the combined State-USAID request (interview with USAID staff, September 10, 2007). There are additional funds in the supplemental requests for Iraq and Afghanistan that might be considered DG programming, but the committee was not able to obtain an estimate for those expenditures.

²For example, in Uganda the work on peace building and reconciliation in Northern Uganda is included in the DG program, and the Peru DG program includes a project to help farmers in coca-producing areas develop alternative crops (see Appendix E for further information).

ment of Transportation for \$12.8 billion simply for highway maintenance and construction (State of California 2007).

The committee stresses at the outset the imbalance between what USAID's missions are asked to do in democracy and governance—to help countries steer their entire participation and governance system in the direction of greater or more stable democracy—and the constrained financial resources they have at their disposal for this task. The committee believes this imbalance is central to any assessment of whether USAID DG projects are actually raising the level of democracy worldwide and also to the way in which the projects are examined to evaluate their impact.

USAID's DG efforts include programs in countries undertaking democratic reforms and countries that are not yet seeking such reforms. Most of the projects are not carried out by USAID personnel but through contracts and grants with private firms and nongovernmental organizations (NGOs). USAID's main role in democracy promotion is thus to plan projects and then select contractors to implement them, or choose local or international NGOs to receive grant support for their activities.

USAID is the single largest provider of funding for democracy assistance. However, in many countries USAID is just one agency among many others providing democracy assistance.³ Although each donor agency plans and carries out its own programs, coordination with other donors occurs on several levels: within countries among donors, through bilateral channels, and through such multilateral venues as the Development Assistance Committee of the Organization for Economic Cooperation and Development (OECD).

USAID's DG Programs

USAID programs to promote DG focus on four distinct but related goals, which are now collectively called "Governing Justly and Demo-

³Some of the other major organizations providing democracy assistance include the United Kingdom's Department for International Development (DfID), Germany's Agency for Technical Cooperation (GTZ), the Swedish Agency for International Development Cooperation (SIDA), the Norwegian Agency for Development Cooperation (NORAD), and the Canadian International Development Agency (CIDA). Many nongovernmental or quasi-governmental organizations, such as the National Endowment for Democracy, the International Foundation for Election Systems, the National Democratic Institute, and the International Republican Institute are also active in international programs of democracy assistance. The Organization for American States is actively promoting democracy in the Americas. Multilateral donor agencies, such as the World Bank, the United Nations Development Program, and the OECD Development Assistance Committee, have also made promotion of good governance (a vague concept but one that overlaps with many elements of democracy, including transparency and accountability of government and impartial rule of law) a priority in their work.

cratically," under the reforms of foreign assistance undertaken by the Bush administration. As shown on the USAID Web site (2007), these are:

- *Strengthening the rule of law and respect for human rights*

The term 'rule of law' embodies the basic principles of equal treatment of all people before the law, fairness, and both constitutional and actual guarantees of basic human rights. A predictable legal system with fair, transparent, and effective judicial institutions is essential to the protection of citizens against the arbitrary use of state authority and lawless acts of both organizations and individuals. . . . Without the rule of law, the executive and legislative branches of government operate without checks and balances, free and fair elections are not possible, and civil society cannot flourish. Beyond the democracy and governance sector, the accomplishment of other USAID goals also relies on effective rule of law.

- *Promoting more genuine and competitive elections and political processes*

Free and fair elections are vital to a functioning democracy. When a country is emerging out of a protracted civil war, or in cases where a country's government has lost the confidence of its citizens, it is often necessary to hold elections very quickly. . . . Competitive political parties are central to any democracy. They perform a number of functions that, in combination, distinguish them from any other civic or social organization.

- *Increased development of a politically active civil society*

The hallmark of a free society is the ability of individuals to associate with like-minded individuals, express their views publicly, openly debate public policy, and petition their government. 'Civil society' is an increasingly accepted term which best describes the nongovernmental, not-for-profit, independent nature of this segment of society.

- *More transparent and accountable governance*

A key determinant for successful democratic consolidation is the ability of democratically-elected governments to provide 'good governance.' . . . 'Good governance' assumes a government's ability to maintain social peace, guarantee law and order, promote or create conditions necessary for economic growth, and ensure a minimum level of social security. Yet many new governments fail to realize the long-term benefits of adopting effective governance policies.

These four goals have remained remarkably constant since the first democracy assistance strategy was adopted in the early 1990s and then enshrined in USAID practice at the outset of the Clinton administration. USAID has thus continued to rely on a consistent framework of challenges and programs to meet them for more than 15 years.

The four broad goals are supported by program components such as Promote Media Freedom, Support Credible Elections, Strengthen Politi-

cal Parties, Strengthen Justice Sector, and Reduce Corruption. In the field these program components are translated into projects, each of which may include many separate activities.⁴ For example, a large stock of projects has been developed to train political parties to compete, to increase civic participation, and to encourage judicial or legislative competence and autonomy. Many DG missions are supporting activities to improve democratic practices within political parties, heighten women's participation in politics, provide technical support to judges or legislators, increase the number of active NGOs, and promote decentralization of government services. As discussed further in the next section, the design and implementation of all of these efforts depend on knowledge and assumptions about what causes, sustains, or hinders the process of democratization.

DEMOCRATIC DEVELOPMENT AND DEMOCRACY ASSISTANCE: WHAT DO WE KNOW?

Ideally, USAID and other providers of DG assistance would be guided in achieving their goals by a well-defined theory of democratic development that could identify where a recipient country stood on feasible trajectories toward stable democracy and which elements or driving factors needed to be supplied or strengthened in order to overcome obstacles and move forward on such a trajectory. It would then select among programs known to provide or strengthen those specific elements and tailor their implementation to that country's specific needs.

Unfortunately, the growth of widely accepted findings regarding the causes and consequences of democratization has lagged behind the growth of democracy assistance activities. Scholars continue to debate exactly how to define democracy, what pathways lead most reliably to full liberal democracy, what the necessary conditions are to achieve and stabilize democracies, and what the consequences are of transitions to democracy for various sets of institutions and geohistorical contexts (Lowenthal 1991, Lijphart 1999, Cox et al 2000, Przeworski et al 2000, Diamond and Plattner 2001, Mansfield and Snyder 2002, Bunce 2003, Chua 2003, Junne and Cross 2003, Acemoglu and Robinson 2005, Pevehouse 2005, Shapiro 2005, Bunce and Wolchik 2006, Tilly 2007). In policy terms this means that scholars can provide only qualified advice on how to move countries

⁴USAID has no standard terms for the various levels of its work. In this report "programs" is used to capture higher levels such as DG, which undertake various "projects" in countries, and these projects in turn may involve multiple "activities." When speaking of evaluating "programs" or "projects" in this report, the committee refers to the evaluation of specific activities to determine whether they are having their desired impact. It is recognized that clusters of such activities may need to be evaluated to assess the overall impact of a large project or, even more broadly, of program activity in a given country or countries.

from dictatorship to stable and full liberal democracy; on how to shore up recently emerged or fragile democracies; or on precisely how to use democratization to address problems of terrorism, domestic or international conflict, or economic decay. It is probably fair to say that scholars know far more about what fully democratic countries look like and how they function than about how nondemocratic or partially democratic countries make the transition to stable full democracies.

These limitations notwithstanding, the field of democracy studies has expanded enormously in the past few decades. In the years immediately following World War II, the main obstacle to the spread of democracy was considered to be communism. Modernization theory argued that if societies could just be kept on a path toward capitalism and free markets, political freedom and democracy would eventually follow.⁵ Yet modernization theory was swept aside in the 1970s and 1980s in a wave of detailed scholarship on the highly varied trajectories of developing, postcolonial, and capitalist and socialist societies. The emergence throughout the developing world in the 1960s and 1970s of a variety of military dictatorships, postcolonial dictatorships, capitalist one-party states, and frequent reversions or collapses of new democratic regimes provoked scholars to reexamine their assumptions. Rather than a nearly inevitable tendency driven by modernization, progress toward democracy came to be seen as a highly problematic process, fragile and prone to reversal.

Building on a few seminal works, from Alexis de Tocqueville's *Democracy in America* to Seymour Martin Lipset's *Political Man* and Robert Dahl's *Polyarchy*, scholars have developed a host of new data and theories regarding democracy. There are at least two journals entirely devoted to democracy studies (*The Journal of Democracy* and *Democratization*), and a multivolume *Encyclopedia of Democracy* (Lipset 1995). The Web site supplement to this report contains a partial bibliography of recent scholarship on democracy and democratization that runs to nearly 20 pages.⁶

This literature falls into three broad groupings. *Cross-national quantitative analyses* seek to identify the average impact of various factors— income, education, culture, religion, or institutional background, for example—on the frequency with which countries undergo democratic transitions or reversions or on the level of democracy as measured by widely used indicators such as the one developed by Freedom House (e.g., Bollen and Jackman 1985, Lewis-Beck and Burhart 1994, Muller and

⁵Modernization theory (Rostow 1960, Huntington 1968) argued that traditional authoritarianism would inevitably give way to demands for mass participation with the spread of industrialization and mass media. Whether such demands gave rise to liberal democracies or communist dictatorships depended on how such mass participation was channeled into politics, whether through competitive party systems or communist one-party states.

⁶See http://www7.nationalacademies.org/dsc/USAID_Democracy_Program.html.

Seligson 1994, Pzeworski et al 2000, Boix and Stokes 2003, Inglehart and Welzel 2005, Epstein et al 2006). *Comparative and historical analyses* seek to identify the key elements in the democratic transitions or outcomes of specific states, usually in a particular region or particular type of transition. Thus, there have been studies of democratization in Latin America, Europe, or Africa and studies of major social revolutions and of peaceful transitions through elite pacts or protest and reforms (e.g., O'Donnell et al 1986; Goldstone et al 1991; Reuschemeyer et al 1992; Linz and Stepan 1996; Bratton and van de Walle 1997; Diamond and Plattner 1998, 1999, 2001; Mahoney 2001; Bunce 2003; Tilly 2004). *Policy research*, which may also include comparative and historical analyses, tends to focus more on policy choices and their consequences and is more likely to try to offer practical advice for decision makers (e.g., Carothers 1999, 2004, 2006b; de Zeeuw and Kumar 2006). *Practitioners' reflections*, a subset of policy research, provide accounts of experiences with programs for democracy promotion or stabilization in various countries, offering "lessons learned" and generalizations to inform other efforts (Dobbins 2003, Durch 2006).

Within each of these groups, controversies and debates have arisen over the definition of democracy and the role of various factors in promoting or consolidating democracy. Moreover, the lack of consensus is as pronounced across as within the various genres. As one eminent scholar has suggested: "We should not search for a single set of circumstances or a repeated series of events that everywhere produces democracy. . . . We should look instead for robust, recurrent causal mechanisms that combine differently, with different aggregate outcomes, in different settings" (Tilly 2004:9).

Rather than providing accepted generalizations on which to base DG programming, the academic literature has been more successful in documenting the great degree of variation in the process of democratization. For example, in the past 50 years, many of the countries that moved toward democracy leapt quickly from dictatorship to democracy (as in Eastern Europe after 1989), while others (such as Mexico, South Korea, and Taiwan) made a series of incremental steps, gradually increasing civil liberties and political competition (Goldstone 2007). There is a clear correlation between higher national incomes and the incidence of stable democracy (Lipset 1960, Barro 1999, Epstein et al 2006), yet a number of relatively poor countries have been successful in sustaining democracy as well (e.g., India, Botswana, Jamaica, and Mauritius). It is also clear that multiple processes have led countries from dictatorship to democracy, ranging from violent revolutions to relatively peaceful protest-driven reforms to pacts orchestrated among elites (e.g., O'Donnell et al 1986, Mahoney 2001, Bunce and Wolchik 2006). Moreover, researchers have not yet concluded that there is a single form of democracy that is most suc-

cessful. Presidential and parliamentary systems, centralized and federal systems, two-party and multiparty systems have all seen both great success and unfortunate failures in diverse countries (Przeworski et al 2000). It is not clear what conclusions should be drawn for democracy assistance from these findings, especially since the academic study of democracy assistance per se, in contrast to studying the broad contours of democracy and democratization, is still in its infancy.⁷

USAID and other democracy assistance agencies therefore face a difficult task. Practitioners' reflections present informed viewpoints and policy research often presents thoughtful and systematic analysis, but their judgments about program success or failure are not rigorously tested according to academic standards. Yet since academic debates regarding democratization remain largely unresolved, they offer little practical guidance on what to do in a given country to build or sustain democracy. Policy professionals working in democracy assistance have therefore formed their own "practical wisdom," based on elements drawn from their readings of the academic and think tank literature, their own experiences, and what they glean from other practitioners. Policy professionals thus often describe democracy promotion as "more of an art than a science," where policy choices must depend on intuition and personal judgment as much or more than on any scientific guidelines.

Despite this range of conflicting findings, there are some things that are known. First, there are more countries that can reasonably claim to be democracies, if only partially achieved, than ever before. Second, among emerging democracies there is considerable variation within and among countries, such that advances are often met with setbacks (Hagopian and Mainwaring 2005). Third, with respect to democracy assistance efforts, one very encouraging finding from recent academic research is that, on average, democracy assistance *does* matter and has a positive impact on democratic progress. Several statistical studies have found that, while controlling for a wide variety of other factors, higher levels of democratic assistance are on average associated with movement from lower

⁷Early and continuing groundbreaking comparative work on the impact of democracy assistance in different contexts was done by Tom Carothers and colleagues at the Carnegie Endowment for International Peace. By contrast, a major center of comparative/historical research on democracy and democratic transition—Stanford University's Center on Democracy, Development, and the Rule of Law (CDDRL)—has only just begun its first studies of the impact of democracy assistance (McFaul 2006). Also, the first statistical analyses of the impact of democracy assistance have only recently begun to appear in major academic journals (Finkel et al 2007, 2008). As further evidence of the relatively immature state of studies of democracy assistance, the Network of Democracy Research Institutes, organized by the National Endowment for Democracy, is just six years old. Harvard University only established its Ash Institute for Democratic Governance in 2003, Stanford University its CDDRL in 2004, and Georgetown University its Center for Democracy and Civil Society in 2002.

to higher levels of democracy, as measured by some of the most general indices of democratic government (Al-Momani 2003; Finkel et al 2007, 2008; Kalyvitis and Vlachaki 2007; Azpuru et al 2008). These effects are robust and statistically significant, providing the clearest evidence to date that democracy assistance generally meets its desired goals.

Thus, despite all of the confusion and conflicting findings, there is a sense that (1) democracy is moving ahead in the world and (2) foreign assistance generally and in some specific cases has made a difference. Unfortunately, it is also true that in a number of highly important cases—such as Haiti, Egypt, and post-Soviet Russia—large volumes of democracy assistance have yielded disappointing results.

It is also alarming that in a number of cases in recent years—Pakistan, Thailand, Bangladesh—countries that seemed on the path to greater democracy have reversed course. There is mounting evidence for a “democracy backlash” in which authoritarian and semiauthoritarian regimes are actively resisting donor efforts as well as internal advocates of democracy (Carothers 2006a). Some research further suggests that authoritarian regimes have become adept at providing economic openings or limited civil liberties to deflect dissent while still maintaining a tight grip on authority (Bueno de Mesquita and Downs 2005). In summary, the conditions that face the United States and USAID for supporting the advance of democracy are growing ever more challenging.

It is therefore crucial, if USAID’s democracy assistance is to be more effective and make best use of scarce resources, that the agency (and other donors) be able to identify which elements of their complex and multifold democracy assistance projects are doing the most work to move democracy forward. Moreover, they would like to know which DG projects work best to accomplish specific goals in particular countries.

USAID’S REQUEST TO THE NATIONAL RESEARCH COUNCIL⁸

Strategic and Operational Research Agenda

USAID has supported external research on many aspects of democracy and governance and undertaken significant internal efforts as part of its search for relevant knowledge and insights to guide its policies. The USAID Web site offers a wide array of publications, covering the range

⁸The National Research Council (NRC) is part of the National Academies, which also includes the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. Created in 1916, the NRC has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities.

from practical manuals filled with lessons learned and best practices to academic research. One of the most significant efforts to determine the impact of democracy assistance began in 2000, when the Office of Democracy and Governance in the Bureau for Democracy, Conflict, and Humanitarian Assistance created the Strategic and Operational Research Agenda (SORA) program. SORA is a long-term effort consisting of a number of research activities. Overall, its goal is to improve the quality of U.S. government DG programs and strategic approaches by (1) analyzing the effectiveness and impact of USAID DG programs since their inception and (2) developing specific findings and recommendations useful to democracy practitioners and policymakers.

The SORA effort has struggled with all of the difficulties of attempting retrospective, comparative analysis of complex foreign policy cases, with a welcome willingness to examine both program successes and failures. The first SORA effort consisted of a set of case studies of democracy programs across six countries, which identified a number of key data and methodological issues for retrospective work (Carter 2001, Carter et al 2003). Another piece, undertaken as a small initial pilot effort, was a "Voices from the Field" project based on in-depth interviews with democracy practitioners to gain their insights about the factors that affected the success or failure of their projects.

In 2003 SORA supported a review of past evaluations of DG programs, which was carried out under the auspices of the Social Science Research Council (SSRC) and later published (Bollen et al 2005). The review found that these evaluations provided useful information for the planning and implementation of DG programs, including insights into management issues, key problems of reception of DG assistance, dealing with local spoilers and other obstacles, and the complexities of carrying out DG programs. These were insightful studies of how DG programs worked, whether or not they worked as expected, and why. However, Bollen et al also found that these past evaluations were not as useful as they hoped for determining the programs' actual effects.

Bollen et al found a serious lack of consistent and systematic information about inputs, activities, and outputs/outcomes in the sample of evaluations they studied. Baseline conditions were rarely fully recorded before implementation of programs began, and no reference or comparison groups were used to help establish whether other trends or external conditions, rather than USAID's DG programs, were responsible for the observed outcomes. They therefore concluded that it would be impossible to use these documents as the basis for a retrospective evaluation of the effectiveness of DG programming.⁹ A subsequent SSRC project recom-

⁹This study and the issues related to USAID's current evaluation process are discussed in greater detail in Chapter 2.

mended a multimethod approach that would both provide for retrospective analysis to learn from past efforts and lay the foundation for a greater capacity to evaluate future DG work (Bollen et al 2003, SSRC 2004).

The first major piece of research sponsored by SORA in response to the SSRC recommendations was the large-scale, cross-national, quantitative analysis examining the effects of USAID democracy assistance on democracy building described above. The first phase of this research, including both the analysis and the dataset, was released in 2006 and the results are beginning to appear in academic journals (Finkel et al 2007, Azpuru et al 2008). The second phase of the quantitative analysis, which added several years of data and examined key issues related to the first set of findings, was released in early 2008 (Finkel et al 2008). The NRC project described in this report is the newest SORA-sponsored activity.

SORA's Charge to the NRC Committee

Noting the problems that the Bollen et al (2005) review found of past USAID project evaluations, USAID asked the NRC for help in developing improved methods for learning about the effectiveness and impact of its work, both retrospectively and in the future. Specifically, the project is to provide:

1. A refined and clear overall research and analytic design that integrates the various research projects under SORA into a coherent whole in order to produce valid and useful findings and recommendations for democracy program improvements.
2. An operational definition of democracy and governance that disaggregates the concept into clearly defined and measurable components.
3. Recommended methodologies to carry out retrospective analysis. The recommendations will include a plan for cross-national case study research to determine program effectiveness and inform strategic planning. USAID will be able to use this plan as the basis of a scope of work to carry out comparative retrospective analysis, allowing USAID to learn from its 25 years of investment in DG programs.
4. Recommended methodologies to carry out program evaluations in the future. The recommendations for future analysis will focus on more rigorous approaches to evaluation than currently used to assess the impact of democracy assistance programming. They should be applicable across the range of DG programs and allow for comparative analysis.
5. An assessment of the feasibility of the final recommended methodologies within the current structure of USAID operations and defining policy, organizational, and operational changes in those operations that might improve the chances for successful implementation.

To respond to USAID's request, the NRC created the ad hoc Committee on the Evaluation of USAID Democracy Assistance Programs, whose members bring expertise in the major areas of USAID activities, direct experience with USAID projects, and expertise in the contributions that social sciences methodology for comparative political analysis could make to improving USAID's evaluations of its work. Appendix A provides biographies of the committee members, and Appendix B gives information about the meetings that were the core of the committee's deliberations.

Responding to the Charge

USAID thus approached the NRC with two broad questions: (1) How can we learn to more effectively support democracy in particular countries and contexts around the world? (2) How can we learn where and whether our specific DG assistance programs have been effective?

While it is tempting for a committee such as this one to draw guidance from current democracy research to advise USAID on how best to pursue democracy assistance in varied circumstances, it is the committee's firm view, based on its review of the evidence, that any such advice would be premature. As already discussed, the current state of the academic literature on democratization is highly contested, and the topic of democracy assistance has only very recently become a focus of academic research. Thus the committee believes that it cannot simply draw on current academic research to answer these questions.

For example, while the committee knows that many researchers would have views on such questions as whether democracy assistance should be "sequenced" in a certain way across sectors, on whether targeting corruption or promoting decentralization are effective ways to advance democracy, and whether democracy assistance is futile in countries under authoritarian rule, it is fairly certain that it would not find widely accepted consensus answers to these questions. Even one of the oldest and most central debates about democracy assistance—whether it is more fruitful to help poor countries develop democracy first, as that will help their subsequent economic growth, or whether economic growth should first be promoted, as this will lay a foundation for subsequent transition to a more lasting democracy—remains far from settled (Przeworski et al 2000, Halperin et al 2004, Gerring et al 2005, Carothers 2007).¹⁰ Moreover, practical problems have raised new issues in democ-

¹⁰The pages of the *Journal of Democracy* are filled with precisely such debates. The chair of this committee observed an event in 2007 at the National Endowment for Democracy that brought together leading researchers on democracy to address a straightforward practical question: What type of voting system (e.g., two-party district-based, party-list proportional,

racy assistance, such as the degree to which democracy assistance can, or should, be pursued in conjunction with security provision by military forces. Past conventional wisdom was that military forces and civilian aid programs should be kept strictly separated, yet conditions in many countries have forced DG programs to work in close partnership with military forces or even for military forces themselves to become agents of government reconstruction and DG assistance (Goldstone 2006, U.S. Department of State 2007). Many questions have arisen about how best to provide democracy assistance in these new circumstances.

Given this uncertainty on broad matters of strategy, the committee has focused on the second question, for there the committee believes it can suggest procedures by which USAID can draw on the work of the academic and policy communities, as well as its own experience in democracy assistance, to make substantial advances in learning which of its DG assistance programs are most effective. Moreover, the committee would go so far as to argue that in the current state of scientific research, answering the second question is likely the best way to also answer the first. That is, the committee believes that the fastest way for USAID to improve the effectiveness of its democracy support programs around the world is to determine which of its programs really work, and how well, in regard to advancing such concrete goals as improving the skills of legislators and the autonomy of judges, reducing corruption, enhancing popular participation, and ensuring free and fair elections. By building its stock of knowledge on which of its DG projects best accomplish these goals, and to what extent and at what cost in specific circumstances, USAID will improve its ability to assist those seeking to advance democracy on the ground in complex and demanding conditions.

At the same time, the committee recognizes that aside from learning about its programs' effectiveness, DG officers require a constant stream of information on program management as well as special evaluations of what happened when things turn out unexpectedly. In addition, since the field of academic research on democracy and democratization is racing ahead, USAID needs to keep abreast of useful findings and be aware of shifts in views and the emergence of consensus when they occur. Thus

or other) would be best suited for particular countries? The experts were wholly unable to agree on anything except that each system has trade-offs and that individual countries would have to choose what they thought met their needs. Even in the most critical cases—such as choosing a voting system for Iraq's first post-Hussein elections—disagreements are severe. The experts who developed the system for Iraq argued that proportional party-list voting would best build on existing organizations' strength and reward high turnout. Other experts argued that such a system would damage democracy by encouraging voting blocs built along ethnic and religious lines. On almost any such aspect of democracy assistance, similar disagreements can be found.

the committee also sees as its responsibility suggesting procedures and organizational reforms that will assist USAID in a broad span of learning activities. These include efforts at improving measures of democracy, learning from comparative and historical case studies of democratization, and developing a diversity of designs for project evaluation. It also includes outlining incentives and procedures to increase active learning and the application of new knowledge and ideas to the planning and implementation of DG activities.

REPORT OVERVIEW

Major Findings and Recommendations

The committee considered both retrospective and prospective approaches to studying USAID activities and how to make best use of methods ranging from case studies to randomized evaluations to the structured sharing of USAID DG officers' experiences through debriefings and conferences. Based on this work, the committee's most important conclusions and recommendations are:

- Most evaluations of DG programs have been designed to meet a variety of diverse monitoring and management needs. While yielding valuable insights, they have not provided compelling evidence of program effects. Collecting the information needed to most clearly determine the impact of DG projects—including before and after measurements on key outcome variables, documentation of changes in policy-relevant outcomes rather than activities completed, and measurements on both the groups receiving assistance and control or comparison groups that did not—is not currently part of most monitoring and evaluation plans for DG programs.

- USAID needs to gain experience with impact evaluations, including those using randomized designs, to learn whether they could improve its ability to more accurately ascertain the effects of its DG programs. If their feasibility is demonstrated for a wide range of DG projects, impact evaluations could provide critical information on what works best, and under what conditions, in democracy assistance.

- Such impact evaluations could take a variety of forms, depending on the character and conditions of specific DG programs. Large N randomized evaluations provide the most accurate and credible determination of the impact of aid programs and should be used where possible. Field studies suggest that many current DG programs (e.g., decentralization programs) could be studied using randomized designs. For those DG programs where randomization is not suitable, other impact evaluation

designs are available, ranging from studies with matched or national baseline comparison groups to single-case studies that use time-series data to examine how outcomes change over time in response to USAID DG assistance.

- There is considerable skepticism, among both scholars and policymakers, regarding the feasibility and appropriateness of applying rigorous impact evaluations to DG activities. On this committee, Larry Garber emphatically shares these concerns. Most of the committee members, on the other hand, while acknowledging and respecting the skepticism among many policymakers, believe that rigorous impact evaluations of DG projects are feasible and that they will provide the most accurate and credible way for U.S. taxpayers as well as the citizens of the countries in which USAID funds democracy programs to gain assurance as to which DG programs work and which do not.

Given these differences in opinion and the need to acquire capacity and experience with using impact evaluations to learn the effects of DG programs, the committee unanimously recommends that USAID begin a pilot initiative designed to demonstrate whether such evaluations can help USAID determine the effects of its DG projects on policy-relevant outcomes. This initiative should include randomized studies and focus on DG projects that are in wide use or represent major investments for USAID; it should also offer expertise and support to missions and DG officers who wish to conduct varied forms of impact evaluations suited to learning about the impact of their programs.

- To better track democratic changes in countries for strategic assessment and policy planning, USAID and other national and international organizations providing democracy assistance should explore making a substantial investment in the systematic collection of democracy indicators at a disaggregated, sectoral level—focused on the components of democracy rather than (or in addition to) the overall concept. Rather than attempting to arrive at a single score capturing all elements of the quality of democracy in a country, this effort should focus on how to best map out a country's politics along a number of discrete dimensions (e.g., civil liberty, transparency, judicial independence, checks on the executive). Such a disaggregated index would allow policymakers to clarify how, specifically, one country's democratic features differ from others in the region or across regions and better identify how changes are occurring over time. These measures should aim to be more transparent, objective, sensitive, and widely accepted than currently available measures of democracy, which have substantial flaws.

- To learn more about the role of its DG assistance projects in a broader range of settings and in varied trajectories of democratization, USAID should either sponsor or seek to gain from ongoing academic

research a more diverse and more theoretically structured set of case studies of democracy assistance than it has developed in the past. The committee suggests that these case studies should examine countries in which USAID has invested substantially and in which it has invested little, countries in which democratization unfolded successfully and where it failed or was reversed, and countries that included a range of varied initial conditions in which DG assistance was offered (e.g., authoritarian or semiauthoritarian regimes, emerging or transitional democracies, and countries emerging from violent internal conflicts).

- To better translate learning into policy planning and effective management, USAID should rebuild its institutional mechanisms for absorbing and disseminating the results of its work and evaluations, as well as its own research and the research of others, on processes of democratization and democracy assistance. This should include conferences, panels, and other creative and active learning opportunities. These should include discussion of its own program evaluations and other research; debate on the work of academics, think tanks, and other donor organizations; and sharing of experiences among DG officers and implementers and other DG assistance providers.

The remainder of this report presents evidence that supports these conclusions and recommendations and offers additional specific recommendations for USAID actions to achieve them.

A Note on Evaluations

Because the main task given the committee by SORA was to provide guidance to USAID on how to determine the effects of its DG programs, this report spends considerable time discussing issues of evaluation design. This is because for the specific task of determining a project's true effects, there is no substitute for a well-designed impact evaluation. Some of this discussion (especially in Chapters 5 and 6) is quite technical because the issues of evaluation design are complicated, especially when dealing with many of the conditions in which USAID must actually work, where USAID does not control the assignment of assistance, conditions are rapidly changing, and pressures from many diverse sources affect programming.

The committee stresses that in its discussion of evaluation practices the committee is not breaking new ground methodologically. If the purpose of an evaluation is to provide evidence that a project has had its intended impact, there is a consensus in the social sciences and program evaluation research communities about the methods that will provide the most confidence in making those judgments (Cook and Campbell 1979, Shadish et al 2001, Wholey et al 2004). Moreover, the committee's recom-

mendations regarding evaluations, and the emphasis on the potential value of undertaking more impact evaluations of aid programs as the best way to improve aid effectiveness, are not unique. Instead, they align with a growing number of recommendations from private foundations, think tanks, and donor governments that have urged greater efforts in exploring the use of impact evaluations to improve DG and other types of foreign assistance.¹¹

It is also recognized, however, that some of the evaluation procedures that the committee (as well as other groups and reports) recommends have not been widely employed in some sectors of the development community, especially in the area of democracy and governance. In fact, as noted above, the committee is aware of significant skepticism among policy professionals and academics regarding the feasibility and appropriateness of applying so-called scientific or randomized evaluation procedures to democracy assistance programming.¹² Perhaps the most important source of skepticism is the belief that applying rigorous impact evaluation procedures to DG programs is impractical given the actual conditions of designing and implementing DG assistance. Committee member Larry Garber strongly noted this point. Or the restrictions on who receives DG programming that is sometimes necessary in order to conduct a rigorous impact evaluation may be considered an unethical failure to respond to an urgent need.

The committee took these objections seriously. What the committee thinks is unique about this project is that we are not drawing on only academic practices or the ideal of how project evaluation should proceed. The committee commissioned fieldwork in three countries—Albania, Peru, and Uganda—to explore the feasibility and desirability of changing evaluation procedures to produce stronger evidence of whether projects were having their intended impact. Independent consultants—chosen for their academic expertise, expertise in the countries or regions visited, and experience in either doing DG-relevant research in the field using the proposed methods or in working with USAID on other aspects of project evaluation—were hired to work with mission DG staff in discussing the potential for revised evaluation procedures.¹³

¹¹These are described in a more detailed discussion of evaluation practices in Chapter 2.

¹²See, for example, the commentary in Banerjee (2007); see also Cook (2006), Davidson (2006), and Scriven (2006). White (2006, 2007) has argued that the portion of development aid that can be subject to randomized impact evaluation is severely limited.

¹³The consultants' biographies can be found in Appendix E, along with the major findings from the visits. The teams were accompanied by a USAID DG staff officer from the SORA project and NRC professional staff members. Following the three field visits, the committee convened a public session at its July 2007 meeting to discuss the findings of the field visits with USAID and a number of DG implementers.

This significantly expanded the range of views and experience available to the committee and, it is hoped, added greater realism to the eventual findings and recommendations. This report uses that field experience to address the most frequently voiced objections regarding the application of more rigorous evaluation procedures to DG programs (see Chapter 6 in particular). In addition, because it is recognized that “best-case” scenarios for employing impact evaluations often cannot be realized, Chapter 7 discusses a large number of “next-best” procedures and practical modifications of DG evaluation practices.

Finally, because only actual experience with using the methods in the field on actual DG projects can truly address the skepticism and concerns about more rigorous evaluations, and because current USAID and implementer capabilities to undertake these methods are limited and would need to be developed, the committee’s actual recommendations are modest and cautious. The committee proposes that a number of impact evaluations, particularly randomized designs, be tested initially through a special initiative aimed at a limited number of thoughtfully chosen DG projects to demonstrate the feasibility and value of such impact evaluations for guiding DG programming.

While this report places great stress on opportunities to build knowledge through exploring the use of impact evaluations, the committee realizes that building knowledge requires more than just efforts to acquire information. The committee therefore recommends that efforts to improve DG project evaluations be part of a broader initiative to restore and augment USAID’s capacity as a learning organization. This initiative should create ongoing programs to involve DG officers throughout the agency in discussion and analysis of research on DG assistance generated inside and outside the agency, including case studies, academic research, and the work of NGOs and other donors. The key to this effort will be the degree to which USAID staff and key implementers are involved in ongoing efforts to share and disseminate their experience, and draw on a variety of sources, to inform program planning and execution.

Plan of the Report

The chapters in this report provide supporting analysis that underpins the committee’s major recommendations. Chapter 2 reviews and assesses current approaches to monitoring and evaluation used by USAID in the context of current evaluation practice in the development assistance community. It distinguishes among various kinds of evaluations for various purposes and discusses how properly designed impact evaluations could make an important addition to USAID’s current mix of monitoring and evaluation practices.

Chapter 3 reviews and analyzes current approaches to measuring democracy and their limitations for USAID's strategic assessment and tracking needs. The analysis draws in part on a workshop held at Boston University in January 2007 to explore the current "state of the art" in the indicators used to track and assess the status of democracy and governance in countries over time (see Appendix C for further information). A somewhat technical chapter, it offers a plan for improving such measures by focusing on measurements at the level of sectoral components of democracy and argues for the need for USAID—either alone or in conjunction with other U.S. government or international agencies—to lead a research project to develop more credible, transparent, objective, and widely accepted measures to track democratic change than current indicators provide. Many of the terms used in this chapter and in Chapters 5 through 7 are defined in the Glossary at the end of this report.

Chapter 4 examines the lessons that can be derived from historical case studies of democratization and democracy assistance and offers suggestions about how USAID can gain more extensive and theoretically structured case studies that would examine the role of democracy assistance in diverse trajectories of democratic development. It draws on a second workshop in March 2007 cosponsored with the Center for Democracy, Development, and the Rule of Law at Stanford University (see Appendix D for further information), which focused on insights that current academic research could provide about democratic transitions and consolidations as a foundation for understanding the potential contributions of democracy assistance.

Chapter 5 returns to the issue of developing sound impact evaluations and provides a theoretical overview of best practices in program evaluation. It examines a variety of designs for impact evaluations, ranging from those suited to projects that involve large numbers of cases with the possibility of randomized assignments to assistance and control groups, to designs where randomization is not possible and for circumstances involving small numbers of cases and even programs with but a single case.

As mentioned above, Chapters 6 and 7 focus on the feasibility of using various evaluation designs to determine the impact of current USAID DG projects, based on lessons from the committee's field visits to DG missions in Albania, Peru, and Uganda. Additional information about the field visits can be found in Appendix E. These chapters explore when randomized assignment might or might not be attainable for actual DG projects, alternatives to randomized assignments, common objections to conducting impact evaluations for DG-type activities, and how to develop impact evaluations in particularly difficult conditions (e.g., one-case situations or cases where USAID has little or no control over which specific

groups or locations receive funding). Chapter 7 also describes how survey research, which is already being widely employed by USAID, could be used for an impact evaluation design, as well as to provide country-level and project-level data for other evaluations.

Chapters 8 and 9 look at USAID's overall organization. Chapter 8 offers proposals for how USAID could adapt its own organizational procedures, either through new efforts or the adjustment of current practices, to reduce the barriers to conducting impact evaluations and, just as important, become more of a "learning organization" that systematically benefits from its own assessments and evaluations and also absorbs lessons from outside researchers and other organizations involved in or studying democracy assistance. Chapter 9 lays out the committee's recommendation for an "evaluation initiative" to test the feasibility of applying impact evaluation methods to DG projects and proposes supporting measures to increase USAID's evaluation capabilities and resources more generally.

REFERENCES

- Acemoglu, D., and Robinson, J.A. 2005. *Economic Origins of Dictatorship and Democracy*. New York: Cambridge University Press.
- Ackerman, S. 2006. Against Democracy. *The American Prospect*. Available at: <http://www.prospect.org/cs/articles?articleId=11933>. Accessed on January 11, 2008.
- Al-Momani, M.H. 2003. Financial Transfer and Its Impact on the Level of Democracy: A Pooled Cross-Sectional Time Series Model. Ph.D. thesis, University of North Texas.
- Azpuru, D., Finkel, S., Pérez-Liñán, A., and Seligson, M.A. 2008. American Democracy Assistance, Patterns and Priorities. *Journal of Democracy* 91(2).
- Banerjee, A.V., ed. 2007. *Making Aid Work*. Cambridge, MA: MIT Press.
- Barro, R.J. 1999. Determinants of Democracy. *Journal of Political Economy* 6:158-183.
- Boix, C., and Stokes, S. 2003. Endogenous Democratization. *World Politics* 55(4):517-549.
- Bollen, K., and Jackman, R.W. 1985. Economic and Noneconomic Determinants of Political Democracy in the 1960s. *Research in Political Sociology* 1:27-48.
- Bollen, K., Paxton, P., and Morishima, R. 2003. Research Design to Evaluate the Impact of USAID Democracy and Governance Programs. Social Science Research Council, New York. Unpublished.
- Bollen, K., Paxton, P., and Morishima, R. 2005. Assessing International Evaluations: An Example from USAID's Democracy and Governance Programs. *American Journal of Evaluation* 26:189-203.
- Bratton, M., and van de Walle, N. 1997. *Democratic Experiments in Africa*. New York: Cambridge University Press.
- Bueno de Mesquita, B., and Downs, G.W. 2005. Democracy and Development. *Foreign Affairs* 84(Sept/Oct).
- Bunce, V. 2003. Rethinking Recent Democratization: Lessons from the Postcommunist Experience. *World Politics* 55(Jan):167-192.
- Bunce, V., and Wolchik, S.L. 2006. Favorable Conditions and Electoral Revolutions. *Journal of Democracy* 17(4):5-18.
- Carothers, T. 1999. *Aiding Democracy Abroad: The Learning Curve*. Washington, DC: Carnegie Endowment.

- Carothers, T. 2004. *Critical Mission: Essays on Democracy Promotion*. Washington, DC: Carnegie Endowment.
- Carothers, T. 2006a. The Backlash Against Democracy Promotion. *Foreign Affairs* 85(Mar/Apr).
- Carothers, T. 2006b. *Confronting the Weakest Link: Aiding Political Parties in New Democracies*. Washington, DC: Carnegie Endowment.
- Carothers, T. 2007. How Democracies Emerge: "The Sequencing" Fallacy. *Journal of Democracy* 18(Jan):12-27.
- Carter, L. 2001. *On the Crest of the Third Wave: Linking USAID Democracy Program Impact to Political Change. A Synthesis of Findings from Three Case Studies*. Washington, DC: Management Systems International.
- Carter, L., Silver, R., and Smith, Z. 2003. *Linking USAID Democracy Program Impact to Political Change: A Synthesis of Findings from Six Case Studies (Draft 2)*. Washington, DC: Management Systems International.
- Chua, A.L. 2003. *The World on Fire: How Exporting Free Market Democracy Breeds Ethnic Hatred and Global Instability*. New York: Anchor Books.
- Congressional Budget Justification: Foreign Operations, Fiscal Year 2008, p. 755. Available at: <http://www.usaid.gov/policy/budget/cbj2008/>. Accessed on September 15, 2007.
- Cook, T.D. 2006. Describing What Is Special About the Role of Experiments in Contemporary Educational Research: Putting the "Gold Standard" Rhetoric into Perspective. *Journal of MultiDisciplinary Evaluation* 6(Nov):1-7.
- Cook, T.D., and Campbell, D.T. 1979. *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.
- Cox, M., Ikenberry J., and Inoguchi, T., eds. 2000. *American Democracy Promotion: Impulses, Strategies and Impacts*. Oxford: Oxford University Press.
- Dahl, R.A. 1971. *Polyarchy: Participation and Opposition*. New Haven, CT: Yale University Press.
- Davidson, J.E. 2006. The RCTs-Only Doctrine: Brakes on the Acquisition of Knowledge? *Journal of MultiDisciplinary Evaluation* 6(Nov):ii-v.
- Diamond, L., and Plattner, M.F., eds. 1998. *Democracy in East Asia*. Baltimore: Johns Hopkins University Press.
- Diamond, L., and Plattner, M.F., eds. 1999. *Democratization in Africa*. Baltimore: Johns Hopkins University Press.
- Diamond, L., and Plattner, M.F. 2001. *The Global Divergence of Democracy*. Baltimore: Johns Hopkins University Press.
- Dobbins, J. 2003. *America's Role in Nation Building: From Germany to Iraq*. Santa Monica, CA: RAND.
- Durch, W.J., ed. 2006. *Twenty-First Century Peace Operations*. Washington, DC: U.S. Institute of Peace.
- Epstein, D., Bates, R., Goldstone, J.A., Kristenson, I., and Halloran, S. 2006. Democratic Transitions. *American Journal of Political Science* 50:551-569.
- Epstein, S., Serafino, N., and Miko, F. 2007. *Democracy Promotion: Cornerstone of U.S. Foreign Policy?* Washington, DC: Congressional Research Service.
- Finkel, S.E., Pérez-Liñán, A., and Seligson, M.A. 2007. The Effects of U.S. Foreign Assistance on Democracy Building, 1990-2003. *World Politics* 59(3):404-439.
- Finkel, S.E., Pérez-Liñán, A., Seligson, M.A., and Tate, C.N. 2008. Deepening Our Understanding of the Effects of U.S. Foreign Assistance on Democracy Building: Final Report. Available at: <http://www.LapopSurveys.org>.
- Gates, R.M. 2007. Landon Lecture. Kansas State University. November 26. Available at: <http://www.defenselink.mil/speeches/speech.aspx?speechid=1199>.
- Gerring, J., Bond, P.J., Barndt, W.T., and Moreno, C. 2005. Democracy and Economic Growth: A Historical Perspective. *World Politics* 57(3):323-364.

- Goldstone, J.A. 2006. The Simultaneity Problem in Stabilization/Reconstruction Operations. *IPOA Quarterly* (Jan):1,10.
- Goldstone, J.A. 2007. Trajectories of Democracy and Development: New Insights from Graphic Analysis. Working Paper, Center for Global Policy, George Mason University.
- Goldstone, J.A., and Ulfelder, J. 2004. How to Construct Stable Democracies. *The Washington Quarterly* 28(1):9-20.
- Goldstone, J.A., Gurr, T.R., and Moshiri, F., eds. 1991. *Revolutions of the Late Twentieth Century*. Boulder, CO: Westview.
- Hagopian, F., and Mainwaring, S., eds. 2005. *The Third Wave of Democratization in Latin America: Advances and Setbacks*. Cambridge: Cambridge University Press.
- Halperin, M.H., Siegle, J.T., and Weinstein, M.M. 2004. *The Democracy Advantage: How Democracies Promote Prosperity and Peace*. New York: Routledge.
- Huntington, S.P. 1968. *Political Order in Changing Societies*. New Haven, CT: Yale University Press.
- Inglehart, R., and Welzel, C. 2005. *Modernization, Cultural Change, and Democracy*. New York: Cambridge University Press.
- Junne, G., and Cross, P. 2003. The Challenge of Postconflict Development. Pp. 1-18 in *Post-Conflict Development: Meeting New Challenges*, G. Junne and W. Verkoren, eds. Boulder: Lynne Rienner.
- Kalyvitis, S.C., and Vlachaki, I. 2007. Democracy Assistance and the Democratization of Recipients. Available at: <http://ssrn.com/abstract=888262>.
- Lewis-Beck, M., and Burkhart, R.E. 1994. Comparative Democracy: The Economic Development Thesis. *American Political Science Review* 88:903-910.
- Lijphart, A. 1999. *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries*. New Haven, CT: Yale University Press.
- Linz, J., and Stepan, A. 1996. *Problems of Democratic Transition and Consolidation*. Baltimore: Johns Hopkins University Press.
- Lipset, S.M. 1960. *Political Man: The Social Bases of Politics*. Garden City, NY: Doubleday.
- Lipset, S.M., ed. 1995. *The Encyclopedia of Democracy*. 4 vols. Washington, DC: Congressional Quarterly.
- Lowenthal, A., ed. 1991. *Exporting Democracy: The United States and Latin America*. Baltimore: Johns Hopkins University Press.
- Mahoney, J. 2001. *The Legacies of Liberalism: Path Dependence and Political Regimes in Central America*. Baltimore: Johns Hopkins University Press.
- Mansfield, E., and Snyder, J. 2002. Democratic Transitions, Institutional Strength, and War. *International Organization* 56(Spring):297-337.
- Mansfield, E., and Snyder, J. 2005. *Electing to Fight: Why Emerging Democracies Go to War*. Cambridge, MA: MIT Press.
- McFaul, M. 2006. *The 2004 Presidential Elections in Ukraine and the Orange Revolution: The Role of U.S. Assistance*. Washington, DC: USAID, Office for Democracy and Governance.
- Muller, E.N., and Seligson, M.A. 1994. Civic Culture and Democracy: The Question of the Causal Relationships. *American Political Science Review* 88:635-654.
- O'Donnell, G., Schmitter, P.C., and Whitehead, L., eds. 1986. *Transitions from Authoritarian Rule*. 4 vols. Baltimore: Johns Hopkins University Press.
- Pevehouse, J. 2005. *Democracy from Above: Regional Organizations and Democratization*. Cambridge: Cambridge University Press.
- Przeworski, A., Alvarez, M.E., Cheibub, J.A., and Limongi, F. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950-1990*. Cambridge: Cambridge University Press.
- Reuschemeyer, D., Stephens, J., and Stephens, E. 1992. *Capitalist Development and Democracy*. Chicago: University of Chicago Press.

- Rostow, W.W. 1960. *The Stages of Economic Growth: A Non-Communist Manifesto*. New York: Cambridge University Press.
- Sanders, P.J., and Halperin, M. 2006. Democracy Promotion as Policy. Council on Foreign Relations. Available at: <http://www.cfr.org/publication/10784/>.
- Scriven, M. 2006. Converting Perspective to Practice. *Journal of Multidisciplinary Evaluation* 6:8-9.
- Shadish, W.R., Cook, T.D., and Campbell, D.T. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, 2nd ed. Boston: Houghton Mifflin.
- Shapiro, I. 2005. *The State of Democratic Theory*. Princeton, NJ: Princeton University Press.
- Social Science Research Council. 2004. Evaluation Plan for USAID Democracy and Governance Activities. Unpublished.
- State of California. 2007. Legislative Analyst's Office (LAO), Analysis of the 2007-2008 Budget Bill, Transportation Chapter. Available at: http://www.lao.ca.gov/analysis_2007/transportation/trans_an107.pdf. Accessed on August 18, 2007.
- Tilly, C. 2004. *Contention and Democracy in Europe, 1650-2000*. Cambridge: Cambridge University Press.
- Tilly, C. 2007. *Democracy*. Cambridge: Cambridge University Press.
- de Tocqueville, A. 1969. *Democracy in America*. 2 vols. Garden City, NY: Anchor Books.
- USAID (U.S. Agency for International Development). 2006. *USAID Primer: What We Do and How We Do It*. Washington, DC: USAID, p. 31. Available at: http://www.usaid.gov/about_usaid/PDACG100.pdf. Accessed on August 7, 2007.
- USAID (U.S. Agency for International Development). 2007. Our Work: Democracy and Governance: Technical Areas. Available at: http://www.usaid.gov/our_work/democracy_and_governance/technical_areas. Accessed on July 10, 2007.
- U.S. Department of State, Bureau of Public Affairs. 2007. Provincial Reconstruction Teams: Building Iraqi Capacity and Accelerating the Transition to Iraqi Self-Reliance. Available at: <http://www.state.gov/documents/organization/78699.pdf>.
- White, H. 2006. *Impact Evaluation—The Experience of the Independent Evaluation Group of the World Bank*. Washington, DC: World Bank.
- White, H. 2007. Technical Rigor Must Not Take Precedence Over Other Kinds of Valuable Lessons. Pp. 81-89 in *Making Aid Work*. A.V. Banerjee, ed. Cambridge, MA: MIT Press.
- Wholey, J.S., Hatry, H.P., and Newcomer, K.E. 2004. *Handbook of Practical Program Evaluation*, 2nd ed. San Francisco: Jossey-Bass.
- de Zeeuw, J., and Kumar, K. 2006. *Promoting Democracy in Postconflict Societies*. Boulder: Lynne Rienner.

Evaluation in USAID DG Programs: Current Practices and Problems

INTRODUCTION

To make decisions about the best ways to assist the spread of democracy and governance (DG), the U.S. Agency for International Development (USAID) must address at least two broad questions:

1. *Where to intervene.* In what countries and in what sectors within countries? Selecting the targets for DG programming requires a theory, or at least hypothesis, about the relationships among different institutions and processes and how they contribute to shaping overall trajectories toward democracy and governance. It also requires strategic assessment, that is, the ability to identify the current quality of democratic institutions and processes in various countries and set reasonable goals for their future development.

2. *How to intervene.* Which DG projects will work best in a given country under current conditions? Learning how well various projects work in specific conditions requires well-designed impact evaluations that can determine how much specific activities contribute to desired outcomes in those conditions.

The two questions are clearly connected. To decide where to intervene (Question 1), one wants to know which interventions can work (Question 2) in the conditions facing particular countries. Indeed, in the current state of scientific knowledge, answers to Question 2 may provide the most helpful guidance to answering Question 1.

This chapter therefore focuses on USAID's policies and practices for monitoring and evaluation (M&E) of its DG projects. To provide a context, we begin with a brief description of the current state of evaluations of development assistance programs in general. Then existing USAID assessment, monitoring, and evaluation practices for DG programs are described. Since such programs are called into existence and bounded by U.S. laws and policies, the key laws and policies that shape current USAID DG assessment and evaluation practices are examined, to lay the foundation for the changes recommended later in the report. The chapter concludes with a discussion of three key problems that USAID encounters in its efforts to decide where and how to intervene.

CURRENT EVALUATION PRACTICES IN DEVELOPMENT ASSISTANCE: GENERAL OBSERVATIONS

As Chapter 5 discusses later in detail, there is a widely recognized set of practices for how to make sound and credible determinations of how well specific programs have worked in a particular place and time (see, e.g., Shadish et al 2001, Wholey et al 2004). The goal of these practices is to determine, not merely what happened following a given assistance program, but how much what happened *differs from what would be observed in the absence of that program*. The final phrase is critical, because many factors other than the given policy intervention—including ongoing long-term trends and influences from other sources—are generally involved in shaping observed outcomes. Without attention to these other factors and some attempt to account for their impact, it is easy to be misled regarding how much an aid program really is contributing to an observed outcome, whether positive or negative.

The practices used to make this determination generally have three parts: (1) collection of baseline data before a program begins, to determine the starting point of the individuals, groups, or communities who will be receiving assistance; (2) collection of data on the relevant desired outcome indicators, to determine conditions after the program has begun or operated for a certain time; and (3) collection of these same “before and after” data for a comparison set of appropriately selected or assigned individuals, groups, or communities that will *not* receive assistance, to estimate what would have happened in the absence of such aid.¹

¹The ideal comparison group is achieved by random assignment, and if full randomization is achieved, a “before” measurement may not be required, as randomization effectively sets the control and intervention groups at the same starting point. However, both because randomization is often not achievable, requiring the use of matched or baseline-adjusted comparison groups, and because baseline data collection itself often yields valuable information about the conditions that policymakers desire to change, we generally keep to the three-part model of sound evaluation design.

Wide recognition of these practices for determining project impacts does not mean that they are widely or consistently applied, however. Nor does it mean that policy professionals or evaluation specialists agree that the three elements are feasible or appropriate in all circumstances, especially for highly diverse and politically sensitive programs such as democracy assistance or other social programs. Thus, while some areas of development assistance, such as public health, have a long history of using impact evaluation designs to assess whether policy interventions have their intended impact, social programs are generally much less likely to employ such methods.

In 2006 the Center for Global Development (CGD), a think tank devoted to improving the effectiveness of foreign assistance in reducing global poverty and inequality, released the report of an “Evaluation Gap Working Group” convened to focus on the problem of improving evaluations in development projects. Their report concludes:

Successful programs to improve health, literacy and learning, and household economic conditions are an essential part of global progress. Yet . . . it is deeply disappointing to recognize that we know relatively little about the net impact of most of these social programs. . . . [This is because] governments, official donors, and other funders do not demand or produce enough impact evaluations and because those that are conducted are often methodologically flawed.

Too few impact evaluations are being carried out. Documentation shows that UN agencies, multilateral development banks, and developing country governments spend substantial sums on evaluations that are useful for monitoring and operational assessments, but do not put sufficient resources into the kinds of studies needed to judge which interventions work under given conditions, what difference they make, and at what cost. (Sayedoff et al 2006:1-2)

Although not a focus for the CGD analysis, democracy assistance reflects this general weakness. As a recent survey of evaluations in democracy programming noted: “Lagging behind our programming, however, is research focusing on the impact of our assistance, knowledge of what types of programming is (most) effective, and how programming design and effectiveness vary with differing conditions” (Green and Kohl 2007:152). The Canadian House of Commons recently investigated Canada’s DG programs and came to similar conclusions:

[W]eaknesses . . . have been identified in evaluating the effectiveness of Canada’s existing democracy assistance funding. . . . Canada should invest more in practical knowledge generation and research on effective democratic development assistance. (House of Commons 2007)

As discussed in more detail below, there are many reasons why DG projects—and social development programs more generally—are not rou-

tinely subject to the highest standards of impact evaluation. One reason is that “evaluation” is a broad concept, of which impact evaluations are but one type (see, e.g., World Bank 2004). On more than one occasion committee members found themselves talking past USAID staff and implementers because they lack a shared vocabulary and understanding of what was meant by “evaluation.”

Diverse Types of Evaluations

Because the term “evaluation” is used so broadly, it may be useful to review the various types of evaluations that may be undertaken to review aid projects.

The type of evaluations most commonly called for in current USAID procedures is *process evaluation*. In these evaluations investigators are chosen after the project has been implemented and spend several weeks visiting the program site to study how the project was implemented, how people reacted, and what outcomes can be observed. Such an evaluation often provides vital information to DG missions, such as whether there were problems with carrying out program plans due to unexpected obstacles, or “spoilers,” or unanticipated events or other actors who became involved. They are the primary source of “lessons learned” and “best practices” intended to inform and assist project managers and implementers. They may reveal factors about the context that were not originally taken into account but that turned out to be vital for program success. Process evaluations focus on “how” and “why” a program unfolded in a particular fashion, and if there were problems, why things did not go as originally planned.

However, such evaluations have a difficult time determining precisely how much any observed changes in key outcomes can be attributed to a foreign assistance project. This is because they often are unable to re-create appropriate baseline data if such data were not gathered before the program started and because they generally do not collect data on appropriate comparison groups, focusing instead on how a given DG project was carried out for its intended participants.

A second type of evaluation is *participatory evaluation*. In these evaluations the individuals, groups, or communities who will receive assistance are involved in the development of project goals, and investigators interview or survey participants after a project was carried out to determine how valuable the activity was to them and whether they were satisfied with the project’s results. Participatory evaluation is an increasingly important part of both process and impact evaluations. In regard to all evaluations, aid agencies have come to recognize that input from participants is vital in defining project goals and understanding what con-

stitutes success for activities that are intended to affect them. This focus on building relationships and engaging people as a project goal means this type of evaluation may also be considered part of regular project activity and not just a tool to assess its effects.

Using participatory evaluations to determine how much a DG activity contributed to democratic progress, or even to more modest and specific goals such as reducing corruption or increasing legislative competence, can pose problems. Participants' views of a project's value may rest on their individual perceptions of personal rewards. This may bias their perception of how much the program has actually changed, as they may be inclined to overestimate the impact of an activity if they benefited from it personally and hope to have it repeated or extended. Thus participatory evaluations should be combined with collection of data on additional indicators of project outcomes to provide a full understanding of project impacts.

Another type of evaluation is an *output evaluation* (generally equivalent to "project monitoring" within USAID). These evaluations consist of efforts to document the degree to which a program has achieved certain targets in its activities. Targets may include spending specific sums on various activities, giving financial support or training to a certain number of nongovernmental organizations (NGOs) or media outlets, training a certain number of judges or legislators, or carrying out activities involving a certain number of villagers or citizens. Output evaluations or monitoring are important for ensuring that activities are carried out as planned and that money is spent for the intended purposes. USAID thus currently spends a great deal of effort on such monitoring, and under the new "F Process," missions report large numbers of output measures to USAID headquarters (more on this below).

Finally, *impact evaluation* is the term generally used for those evaluations that aim to establish, with maximum credibility, the effects of policy interventions relative to what would be observed in the absence of such interventions. These require the three parts noted above: collection of baseline data; collection of appropriate outcome data; and collection of the same data for comparable individuals, groups, or communities that, whether by assignment or for other reasons, did and did not receive the intervention.

The most credible and accurate form of impact evaluation uses randomized assignments to create a comparison group; where feasible this is the best procedure to gain knowledge regarding the effects of assistance projects. However, a number of additional designs for impact evaluations exist, and while they offer somewhat less confidence in inferences about program effects than randomized designs, they have the virtue of being applicable in conditions when randomization cannot be applied

(e.g., when aid goes to a single group or institution or to a small number of units where the donor has little or no control over selecting who will receive assistance).

Impact evaluations pose challenges to design, requiring skill and not merely science to identify and collect data from an appropriate comparison group and match the best possible design to the conditions of the particular assistance program. The need for baseline data on both the group receiving the policy intervention and the comparison group usually means that the evaluation procedures must be designed before the project is begun and carried out as the project itself is implemented. Finally, the need to collect baseline data and comparison group data may increase the costs of evaluation.

For these reasons, among others, impact evaluations of DG programs are at present the most rarely carried out of the various kinds of evaluations described here. Indeed, many individuals throughout the community of democracy assistance donors and scholars have doubts about the feasibility and utility of conducting rigorous impact evaluations of DG projects. Within the committee, Larry Garber has strongly expressed concerns in this regard, and the committee as a whole has given a great deal of attention to these worries. However, as discussed in Chapters 6 and 7, there are a number of practical ways to deal with these issues, and these were explored in the field by the committee's consultants in partnership with several missions. In addition, a good evaluation design is not necessarily more expensive or time-consuming than routine monitoring or a detailed process evaluation.

The differences among these distinct kinds of evaluations are often obscured by the way in which the term "evaluation" is used in DG and foreign assistance discussions. "Evaluation" is often used to imply any estimate or appraisal of the effects of donor activities, ranging from detailed counts of participants in specific programs to efforts to model the aggregate impact of all DG activities in a country on that country's overall level of democracy. This catch-all use of the term "evaluation" undermines consideration of whether there is a proper balance among various kinds of evaluations, how various types of evaluations are being used, and whether specific types of evaluations are being done or are needed. As another CGD report notes:

Part of the difficulty in debating the evaluation function in donor institutions is that a number of different tasks are implicitly simultaneously assigned to evaluation: building knowledge on processes and situations in receiving countries, promoting and monitoring quality, informing judgment on performance, and, increasingly, measuring actual impacts. Agencies still need their own evaluation teams, as important knowledge providers from their own perspective and as contributors to quality

management. But these teams provide little insight into our actual impacts and, although crucial, their contribution to knowledge essentially focuses on a better understanding of operational constraints and local institutional and social contexts. All these dimensions of evaluations are complementary. For effectiveness and efficiency reasons, they should be carefully identified and organized separately: some need to be conducted in house, some outside in a cooperative, peer review, or independent manner. In short, evaluation units are supposed to kill all these birds with one stone, while all of them deserve specific approaches and methods. (Jacquet 2006)

Efforts to Improve Assessments and Evaluations by Donor Agencies

There are encouraging signs of efforts to put greater emphasis on impact evaluations for improving democracy and governance programs. The basic questions motivating USAID's Strategic and Operational Research Agenda (SORA) project are also motivating other international assistance agencies and organizations. The desire to understand "what works and what doesn't and why" in an effort to make more effective policy decisions and to be more accountable to taxpayers and stakeholders has led a host of agencies to consider new ways to determine the effects of foreign assistance projects.

This focus on impact evaluations in particular has increased since the creation of the Millennium Challenge Corporation (MCC) and the 2005 Paris Declaration on AID Effectiveness. Yet while there is wide agreement that donors need more knowledge of the effects of their assistance projects, and there are increased efforts to coordinate and harmonize the approaches and criteria employed in pursuit of that knowledge, donors are far from consensus on how best to answer the fundamental questions at issue. As the Organization for Economic Cooperation and Development (OECD) has stated:

There is strong interest among donors, NGOs and research institutions in deepening understanding of the political and institutional factors that shape development outcomes. All donors are feeling their way on how to proceed. (OECD 2005:1)

Several donors have focused on the first question posed above, the question of where to intervene in the process of democratization to help further that process. In the committee's view this is a question that the current state of knowledge on democratic development cannot answer. It is an essential question, however, and Chapters 3 and 4 suggest specific research programs that might help bring us closer to answers. These issues are more a matter of strategic assessment of a country's condition and potential for democratic development, rather than evaluation, a term

the committee thinks is better reserved for studying the effects of specific DG programs. Nonetheless, several national development assistance agencies have, under the general rubric of improving evaluation, sought to improve their strategic assessment tools. What all of the following donor programs have in common is an increased effort at acquiring and disseminating knowledge about how development aid works in varied contexts.

The broad range of current efforts to revise and improve evaluation procedures undertaken by national and international assistance agencies described below are aimed at better understanding the fundamental questions of interest to all: “what works and what doesn’t and why,” although at present only some involve the use of impact evaluations.

Perhaps the most visible leader in efforts to increase the use of impact evaluations is MCC, which has set a high standard for the integration of impact evaluation principles into the design of programs at the earliest stages and for the effective use of baseline data and control groups:

There are several methods for conducting impact evaluations, with the use of random assignment to create treatment and control groups producing the most rigorous results. Using random assignment, the control group will have—on average—the same characteristics as the treatment group. Thus, the only difference between the two groups is the program, which allows evaluators to measure program impact and attribute the results to the MCC program. For this reason, random assignment is a preferred impact evaluation methodology. Because random assignment is not always feasible, MCC may also use other methods that try to estimate results using a credible comparison group, such as double difference, regression discontinuity, propensity score matching, or other type of regression analysis. (MCC 2007:19)

The World Bank has also embarked on the use of impact evaluations for aid programs through its Development Impact Evaluation (DIME) project. Many of the DIME studies involve randomized-experimental evaluations; moreover, “rather than drawing policy conclusions from one-time experiments, DIME evaluates portfolios of similar programs in multiple countries to allow more robust assessments of what works” (Banerjee 2007:30).²

A major symposium on economic development aid also recently explored the pros and cons of conducting impact evaluations of specific programs (Banerjee 2007). While there were numerous objections to the unrestrained use of such methods (which are explored in more detail in Chapters 6 and 7 below), many eminent contributors urged that foreign

²The CGD has also created the International Initiative for Impact Evaluation to encourage greater use of this method. See http://www.cgdev.org/section/initiatives/_active/evalgap/calltoaction.

aid cannot become more effective if we are unwilling to subject our assumptions about how well various assistance programs work to credible tests. The lead author argued that ignorance of general principles to guide successful economic development (a situation that applies as much or more to our knowledge of democratization) is a powerful reason to take the more humble step of simply trying to determine which aid projects in fact work best in attaining their specific goals.

The Department for International Development (DfID) of the United Kingdom has developed the “Drivers of Change” approach because “donors are good at identifying *what* needs to be done to improve the lives of the poor in developing countries. But they are not always clear about how to make this happen most effectively” (DfID 2004:1). By focusing on the incorporation of “underlying political systems and the mechanics of pro-poor change . . . in particular the role of institutions—both formal and informal” into their analysis, this approach attempts to uncover more clearly what fosters change and reduces poverty. This approach is currently being widely applied to multiple development contexts and is being taught to numerous DfID country offices (OECD 2005:1).

Multipronged approaches to evaluation are being employed by the German Agency for Technical Cooperation (Deutsche Gesellschaft für Technische Zusammenarbeit, GTZ). The range of instruments currently being employed is based on elements of self-evaluation as well as independent and external evaluations. Evaluations aim to address questions of relevance, effectiveness, impact, efficiency, and sustainability.³ These questions are addressed throughout the project’s life span as a means of better understanding the links between inputs and outcome. Commitment by the GTZ to evaluations is demonstrated by the agency’s increased spending on these activities, spending “roughly 1.2 percent of its public benefit turnover on worldwide evaluations—some EUR 9 million a year” (Schmid 2007).

The Swedish Agency for International Development Cooperation (SIDA) is also actively considering ways to improve its evaluation tools. Since 2005, SIDA has shifted from post-hoc project evaluations to a focus on underlying assumptions and theories; specifically, SIDA is currently conducting a project that “looks at the program theory of a number of different projects in the area. This evaluation focuses on the theoretical constructs that underpin these projects and tries to discern patterns of

³For further information, see “Working on Sustainable Results: Evaluation at GTZ.” Available at: <http://www.gtz.de/en/leistungsangebote/6332.htm>. Accessed on September 12, 2007.

ideas and assumptions that recur across projects and contexts.”⁴ Building on these initial efforts, SIDA hopes to combine the results of this study with others to “make an overall assessment of the field.”

The Norwegian Agency for Development Cooperation (NORAD) has also initiated a new strategy for evaluating the effectiveness of its programs in the area of development assistance. The intent of this new strategy, undertaken in 2006, is to “help Norwegian aid administrators learn from experience by systematizing knowledge, whether it is developed by (themselves), in conjunction with others, or entirely by others. Additionally, the evaluation work has a control function to assess the quality of the development cooperation and determine whether resources applied are commensurate with results achieved.”⁵ Additional attention is being paid to communicating the results of such evaluations with other agencies and stakeholders; this emphasis on communicating results is widely shared in the donor community.

The Danish Ministry of Foreign Affairs has embarked on an extensive study of both its own and multilateral agencies’ evaluations of development and democracy assistance (Danish Ministry of Foreign Affairs 2005). It has found that evaluations vary greatly in method and value, with many evaluations failing to provide unambiguous determinations of program results. In regard to the United Nations Development Program’s central evaluation office, “its potential for helping strengthen accountability and performance assessment is being underexploited, both for the purpose of accountability and as an essential basis for learning” (Danish Ministry of Foreign Affairs 2005:4).

Finally, the Canadian International Development Agency (CIDA) has been involved in recent efforts to improve evaluation and learning from collective experiences at international assistance in the area of democracy and governance.

In April 1996, as part of its commitment to becoming more results-oriented, CIDA’s President issued the “Results-Based Management in CIDA—Policy Statement.” This statement consolidated the agency’s experience in implementing Results-Based Management (RBM) and established some of the key terms, basic concepts and implementation principles. It has since served as the basis for the development of a variety of management tools, frameworks, and training programs. The Agency Accountability Framework, approved in July 1998, is another

⁴For more information on this project, see SIDA, “Sida’s Work with Democracy and Human Rights.” Available at: http://www.sida.se/sida/jsp/sida.jsp?d=1509&a=32056&language=en_US. Accessed on September 12, 2007.

⁵For more information, see NORAD’s Web site: http://www.norad.no/default.asp?V_ITEM_ID=5704. The new strategy discussed here can be found at <http://www.norad.no/items/5704/38/7418198779/EvaluationPolicy2006-2010.pdf>. Accessed on September 12, 2007.

key component of the results-based management approach practiced in CIDA. (CIDA 2007)

The CIDA report makes an important distinction, however: “The framework articulates CIDA’s accountabilities in terms of developmental results and operational results at the overall agency level, as well as for its various development initiatives. This distinction is crucial . . . since the former is defined in terms of actual changes achieved in human development through CIDA’s development initiatives, while the latter represents the administration and management of allocated resources (organisational, human, intellectual, physical/material, etc.) aimed at achieving development results.”

In short, there is growing agreement—across think tanks, blue-ribbon panels, donor agencies, and foreign ministries—that current evaluation practices in the area of foreign assistance in general, and of democracy assistance in particular, are inadequate to guide policy and that substantial efforts are needed to improve the knowledge base for policy planning. Thus, USAID is not alone in struggling with these issues.

CURRENT POLICY AND LEGAL FRAMEWORK FOR USAID DG ASSESSMENTS AND EVALUATIONS

Current DG policies regarding project assessment and evaluation are shaped in large part by broader USAID and U.S. government policies and regulations. Official USAID policies and procedures are set forth in the Automated Directives System (ADS) on its Web site; Series 200 on “Programming Policy” covers monitoring and evaluation in Section 203 on “Assessing and Learning” (USAID ADS 2007). Of particular importance for this report, in 1995 the USAID leadership decided to eliminate the requirement of a formal evaluation for every major project; instead evaluations would be “driven primarily by management need” (Clapp-Wincek and Blue 2001:1). The prior practice of conducting mainly post-hoc evaluations (which were almost entirely process evaluations), often done by teams of consultants brought in specifically for the task, was seen as too expensive and time consuming to be applied to every project.

As a result of the change, the number of evaluations for *all types* of USAID assistance, not just DG, has declined, and the approach to evaluation has evolved over time (Clapp-Wincek and Blue 2001). ADS 203.3.6.1 (“When Is an Evaluation Appropriate?”) lists a number of situations that should require an evaluation:

- A key management decision is required, and there is inadequate information;
- Performance information indicates an unexpected result (posi-

tive or negative) that should be explained (such as gender differential results);

- Customer, partner, or other informed feedback suggests that there are implementation problems, unmet needs, or unintended consequences or impacts;
 - Issues of sustainability, cost effectiveness, or relevance arise;
 - The validity of Results Framework hypotheses or critical assumptions is questioned (e.g., due to unanticipated changes in the host country environment);
 - Periodic Portfolio Reviews have identified key questions that need to be answered or that need consensus; or
 - Extracting lessons is important for the benefit of other Operating Units or future programming (USAID ADS 2007:24).

These evaluations generally remain the traditional process evaluations using teams of outside experts undertaken while a project is under way or after it has been completed.⁶

The second significant policy shaping USAID evaluation practices is the Government Performance and Results Act (GPRA) of 1993. GPRA “establishes three types of ongoing planning, evaluation, and reporting requirements for executive branch agencies: strategic plans . . . , annual performance plans, and annual reports on program performance. In complying with GPRA, agencies must set goals, devise performance measures, and then assess results achieved” (McMurtry 2005:1). GPRA has led to the development of an elaborate performance monitoring system across the federal government. Performance monitoring is different from evaluation; as defined by USAID, for example:

Performance monitoring systems track and alert management as to whether actual results are being achieved as planned. They are built around a hierarchy of objectives logically linking USAID activities and resources to intermediate results and strategic objectives through cause-and-effect relationships. For each objective, one or more indicators are selected to measure performance against explicit targets (planned results to be achieved by specific dates). Performance monitoring is an ongoing, routine effort requiring data gathering, analysis, and reporting on results at periodic intervals.

Evaluations are systematic analytical efforts that are planned and conducted in response to specific management questions about performance of USAID-funded development assistance programs or activities. Unlike

⁶Clapp-Wincek and Blue (2001), for example, define evaluation as “any empirically-based analysis of problems, progress, achievement of objectives or goals, and/or unintended consequences for missions” (p. 2).

performance monitoring, which is ongoing, evaluations are occasional—conducted when needed. Evaluations often focus on why results are or are not being achieved. Or they may address issues such as relevance, effectiveness, efficiency, impact, or sustainability. Often, evaluations provide management with lessons and recommendations for adjustments in program strategies or activities. (USAID 1997:1)

To implement the system required by GPRA, every USAID operating unit (missions overseas, bureaus or offices in Washington) must develop strategic objectives (SOs). The DG office created a process for strategic assessments that is often used to inform the development of mission strategies (USAID 2000). Typically, a team of experts, which may include a mix of contractors and USAID personnel, spends several weeks evaluating current conditions in a country with respect to key aspects of democracy and governance and analyzing the opportunities for intervention and impact. This assessment is not quite keyed to the four elements in USAID's DG goals described in Chapter 1, however. Rather, strategic assessments deal with five areas: consensus, rule of law, competition, inclusion, and good governance. After surveying the degree to which the country has these elements, the assessment considers the key actors and institutions whose behavior or condition needs to change to improve democratic development and then suggests policies—with explicit attention to feasibility given the resources of USAID in that country and country conditions—to promote advances in some or all of these areas. Not every country is assessed and some country assessments may be updated if conditions change enough to warrant a reexamination. Since the formal assessment tool was adopted in late 2000, more than 70 assessments have been conducted in 59 countries.⁷

To achieve their strategic objectives, all USAID operating units develop a Results Framework and a Program Monitoring Plan that include subobjectives that tie more closely to specific projects (see Figure 2-1 for an illustrative results framework). Depending on the size of its budget and other factors, a mission might have anywhere between one and a dozen SOs, of which one or perhaps two will relate to democracy and governance.

Indicators are used to track progress from the project level through intermediate objectives up to the SO. Missions are required to report their performance against these indicators annually, but below the SO level they can choose which indicators to report and can change the indicators they report each year. Generally, each contract or grant must have an

⁷Interviews with USAID personnel, August 1, 2007 and March 3, 2008. Not all of the assessments are made public because missions sometimes consider the judgments politically sensitive.

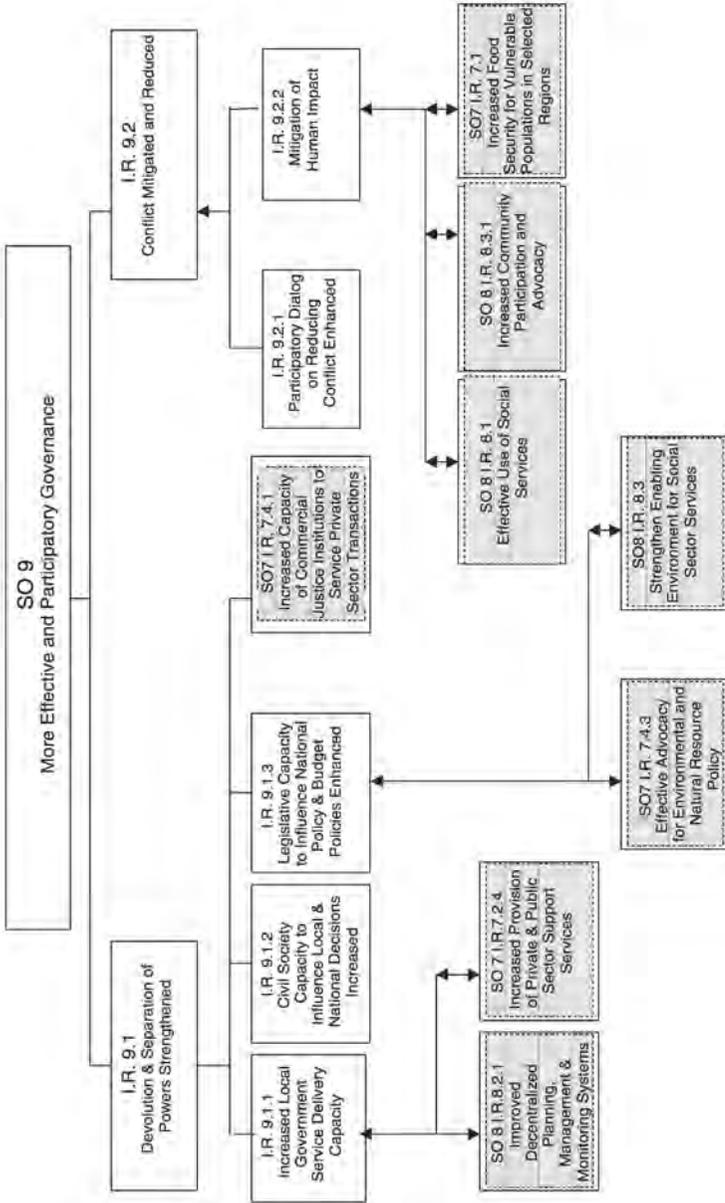


FIGURE 2-1 Illustrative Results Framework.
 SOURCE: USAID 2004. Integrated Strategic Plan for USAID's Program in Uganda, 2002-2007. vol 1: Assisting Uganda to reduce mass poverty. Washington, DC: USAID.

approved performance monitoring plan, which includes both targets and the indicators that will be used to determine whether the project meets its objectives (USAID ADS 2007). Some implementers also develop and track additional indicators, usually to provide further evidence of achieving project goals.

In DG alone, thousands of indicators are used every year to track project performance. Most of them are related to the outputs of specific activities or very proximate project outcomes. This process, supplemented by occasional evaluations, constitutes the largest portion of what USAID refers to as “monitoring and evaluation.” The results of this process are that USAID DG missions spend a large amount of time and money acquiring and transmitting the most basic accounting-type information on their projects (what is described above as “output” evaluations); far less time and money are spent in determining which projects really work and how efficient they are at producing desired results.

In January 2006, Secretary of State Condoleezza Rice initiated a series of reforms, centered on the budget and program planning process, intended to bring greater coherence to U.S. foreign assistance programs (USAID 2006). As part of these reforms the USAID administrator was designated the director of foreign assistance (DFA) and provided with a staff in the State Department to supplement the staff of USAID in implementing the reforms. Instead of a largely bottom-up process that collected, coordinated, and eventually reconciled budget and program requests from individual offices and missions, the new F Process exercised an unprecedented degree of centralized control, setting common objectives for State and USAID and bringing most budget and programming decisions to Washington.⁸ Eventually, the first joint State-USAID budget was submitted to Congress for FY2008, with significant changes in aid allocations for a number of countries (Kessler 2007).

Creation of the DFA structure in the State Department led to the dissolution of the separate policy planning apparatus in USAID. As part of this change, the Center for Development Information and Evaluation (CDIE), which served as a clearinghouse for all evaluations in USAID and had also commissioned the series of independent evaluations of USAID DG programs discussed above, was dissolved and its personnel were transferred into the new DFA Office of Strategic Information in the State Department.

The F Process also resulted in the creation of a set of common indicators collected for all programs in all missions. Most of these are output measures, which for the first time provided a comprehensive look at

⁸A number of projects, however, including the MCC and the President’s Emergency Fund for AIDS, were not included in the F Process for the FY2008 budget.

USAID activities worldwide (U.S. Department of State 2006). Their use in DG is examined in greater detail below. While these output indicators are designed to reflect the overall level of USAID DG activity in a country, they are not intended to provide a strategic assessment of levels of democracy in a country or evidence of the impact of specific DG projects.

Any recommendations for changing the approach to evaluation of DG programs will have to operate within this broader context in USAID and the wider donor community. Within USAID the GPRA-required structure of SOs for programs and performance monitoring for projects is a legal mandate that USAID can adapt but not eliminate. How much of the F Process will endure is unclear at present, but it does illustrate how much can happen—and how quickly—with high-level leadership and support.

THREE KEY PROBLEMS WITH CURRENT USAID MONITORING AND EVALUATION PRACTICES

Focusing on Appropriate Measures Regarding DG Activities

As noted above, USAID has developed many good indicators to track the results of its DG projects. USAID is clearly aware of the important differences between various levels of indicators—those dealing with attaining targeted outputs, those dealing with the institutional or behavioral changes sought by the program, those dealing with broad sectoral changes at the country level, and those dealing with national levels of democracy. The *Handbook of Democracy and Governance Program Indicators*, developed by the Center for Democracy and Governance (USAID 1998) as part of the implementation of GPRA, is the most comprehensive collection of indicators in this area of which the committee is aware. It sets forth detailed suggestions on how to measure outputs and outcomes in the four areas of concern to the DG office: rule of law, elections and political processes, civil society, and governance. It provides a valuable resource to missions and subcontractors as they develop appropriate indicators to assess the impact of specific programs in these sectors.

The development of output measures, especially in some program areas, has continued. The following is taken from the draft of a handbook on support for decentralization programming, currently being prepared for use by USAID:

A distinction should be drawn at the outset between two different kinds of M&E [monitoring and evaluation] activities. One kind of M&E seeks to assess progress on program implementation, that is, the process of implementing decentralization reforms. To this end, one might gather and analyze data on what are sometimes called output indicators: the number of meetings and workshops held, officials trained, and so on.

These kinds of indicators can help to document whether necessary steps are being taken towards effective support of decentralization programs, and they may be especially useful as management tools for program implementation.

Another kind of M&E, however, seeks to assess the impact of decentralization programming on the broader goals described in this handbook: enhancing stability, promoting democracy, and fostering economic development. The key questions are whether and how we can attribute outcomes along these dimensions, or aspects of these dimensions, to the effect of USAID initiatives in support of decentralization programming. This kind of M&E is crucial, for it is the only way to assess what works and what does not in decentralization programming. (USAID 2007)

A few of the democracy indicators recommended by this handbook include:

- Ease with which political parties can register to participate in elections;
- Ability of independent candidates to run for office;
- Number of human rights violations, as tracked by civil society organizations (CSOs) or ombudsman's office;
- Proportion of citizens who positively evaluate government responsiveness to their demands;
- Existence of competitive local elections;
- Percentage of total subnational budget under the control of participatory bodies.

USAID has also funded various agencies to collect valuable data on outcome indicators. For example, a recent national survey in Afghanistan conducted by the Asia Foundation (2007) and underwritten by USAID collected data on the following indicators and many others:

- Do you agree or disagree with the statement that some people make: "I don't think the government cares much about what people like me think."
- How would you rate the security situation in your area: Excellent, good, fair, or poor?
- Compared to a year ago, do you think the amount of corruption overall in your neighborhood has increased, stayed the same, or decreased? In your province? In Afghanistan as a whole?
- Would you participate in the following activities with no fear, some fear, or a lot of fear: voting, participating in a peaceful demonstration, running for public office?

Such survey questions make excellent baseline indicators on outcome measures for many DG assistance projects. USAID could then survey assisted and nonassisted groups on the same questions a year later to help determine the impact of DG assistance. This is an example where USAID can make use of extant surveys that already provide baseline data on a variety of relevant outcome measures.

A more centralized set of indicators was developed as part of the F Process. As mentioned above, the Foreign Assistance Performance Indicators are intended to measure “both what is being accomplished with U.S. foreign assistance funds and the collective impact of foreign and host-government efforts to advance country development” (U.S. Department of State 2006). Indicators are divided into three levels: (1) the Objective level, which are usually country-level outcomes, as collected by other agencies such as the World Bank, United Nations Development Program, and Freedom House; (2) the Area level, measuring performance of sub-sectors such as “governing justly and democratically,” which captures most of the objectives pursued by the DG office; and (3) the Element level, which seeks to measure outcomes that are directly attributable to USAID programs, projects, and activities, using data collected primarily by USAID partners in the field (U.S. Department of State 2006).

Clearly, USAID has taken the task of performance-based policymaking seriously. The central DG office, the various missions throughout the world, and the implementers who support USAID’s work in the field are all acutely aware of the importance of measurement and the various obstacles encountered. The concerns the committee heard were often not that USAID lacks the right measures to track the outcomes of its programs. Although this can be a major problem for some areas of DG, the committee also saw evidence that USAID field missions and implementers have, and seek to use, appropriate measures for program outcomes. Rather, the problem is that the demands to supply detailed data on basic output measures or to show progress on more general national-level measures overwhelm or sidetrack efforts that might go into collecting data on the substantive outcomes of projects.

Matching Tasks with Appropriate Measurement Tools

Broadly speaking, USAID is concerned with three measurement-related tasks: (1) project monitoring, (2) project evaluation, and (3) country assessment. The first concerns routine oversight (e.g., whether funds are being properly allocated and implementers are adhering to the terms of a contract). The second concerns whether the program is having its intended effect on society. The third concerns whether a given country

is progressing or regressing in a particular policy area with regard to democratization (USAID 2000).

Corresponding to these different tasks are three basic types of indicators: *outputs*, *outcomes*, and *meso- and macro-level indicators*. Output measures track the specific activities of a project, such as the number of individuals trained or the organizations receiving assistance. Outcome measures track policy-relevant factors that are expected to flow from a particular project (e.g., a reduction in corruption in a specific agency, an increase in the autonomy and effectiveness of specific courts, an improvement in the fairness and accuracy of election vote counts). Meso- and macro-level measures are constructed to assess country-level features of specific policy areas and are often at levels of abstraction that are particularly difficult to determine with any exactness. Examples include “judicial autonomy,” “quality of elections,” “strength of civil society,” and “degree of political liberties.” For purposes of clarification, these concepts are included, along with an illustrative example, in Table 2-1.

As noted, USAID has made extensive efforts to identify indicators at all levels and across a wide range of sectors of democratic institutions. Nonetheless, in practice a mismatch often arises between the chosen measurement tools and the tasks these tools are expected to perform. Two problems, in particular, stand out. First, based on the committee’s discussions with USAID staff and implementers and further discussions and reviews of project documents during the three field visits described

TABLE 2-1 Measurement Tools and Their Uses

	1. Output	2. Outcome	3. Meso-Level Indicator	4. Macro-Level Indicator
Definition	Indicator focused on counting activities or immediate results of a program	Indicator focused on policy-relevant impacts of a program	Indicator focused on broad national characteristics of a policy area or sector	Indicator focused on national levels of democracy
Level	Generally subnational	Generally subnational	National	National
Example: Improving elections	Number of polling stations with election observers	Reduction in irregularities at the polls (bribing, intimidation)	Quality of election	Level of democracy (e.g., Freedom House Index of Political Rights)
Objective	Monitoring	Evaluation	Assessment	Assessment

in Chapter 7, there is continuing concern that the effectiveness of specific USAID DG projects should not be judged on the basis of meso- or macro-level indicators, such as the overall quality of elections or even changes in national-level indicators of democracy. Second is whether current practices lead to overinvestment in generating and collecting basic output measures, as opposed to policy-relevant indicators of project results.

The F Process indicators reflect both of these problems, although they had little impact on day-to-day project implementation during the course of this study. As noted above, these mandate collecting data at the “Objective” and “Area” levels, which correspond to macro- and meso-level indicators in the table, and at the “Element” level, which corresponds mostly to the output level. Data at the outcome level, which seems crucial to evaluating how well specific projects actually achieve their immediate goals, thus suffer relative neglect.

USAID mission staff and program implementers complained that the success of their projects was being judged (in part) on the basis of macro-level indicators that bore very little or no plausible connection to the projects they were running, given the limited funds expended and the macro nature of the indicator. The most common example given was the use of changes in the Freedom House Political Rights or Civil Liberties Index as evidence of the effectiveness or ineffectiveness of their projects, even though these national-level indices were often quite evidently beyond their control to affect. One implementer commented that his group had benefited from an apparent perception that his project had contributed to improvements in the country’s Freedom House scores over the past several years. While this coincidence worked in his firm’s favor, he made it clear that this was purely coincidental; he was also concerned that if the government policies that currently helped his work changed and made his work more difficult, this would be taken as evidence that his project had “failed.”

This is a poor way to measure project effectiveness. To use the example in Table 2-1, although USAID may contribute to better elections or even more democracy in a nation as a whole, there are always multiple forces and often multiple donors at work pursuing these broad goals. USAID may be very successful in helping a country train and deploy election monitors and thus reduce irregularities at the polling stations. But if the national leaders have already excluded viable opposition candidates from running, or deprived them of media access, the resulting flawed elections should not mean that USAID’s specific election project was not effective. As a senior USAID official with extensive experience in many areas of foreign assistance has written regarding this problem:

To what degree should a specific democracy project, or even an entire USAID democracy and governance programme, be expected to have an

independent, measurable impact on the overall democratic development in a country? Th[at] sets a high and perhaps unreasonable standard of success. Decades ago, USAID stopped measuring the success of its economic development programmes against changes in the recipient countries' gross domestic product (GDP). Rather, we look for middle-level indicators: we measure our anti-malaria programmes in the health sector against changes in malaria statistics, our support for legume research against changes in agricultural productivity. What seems to be lacking in democracy and governance programmes, as opposed to these areas of development, is a set of middle-level indicators that have two characteristics: (a) we can agree that they are linked to important characteristics of democracy; and (b) we can plausibly attribute a change in those indicators to a USAID democracy and governance programme. It seems clear that we need to develop a methodology that is able to detect a reasonable, plausible relationship between particular democracy activities and processes of democratic change. (Sarles 2007:52)

The appropriate standard for evaluating the effectiveness of specific DG projects and even broader programs is how much of the targeted improvement in behavior and institutions can be observed *compared to conditions in groups not supported by such projects or programs*. It is in identifying how much difference specific programs or projects made, relative to the investment in such programs, that USAID can learn what works best in given conditions.

Of course, it is hoped that such projects do contribute to broader processes of democracy building. But these broader processes are subject to so many varied forces—from strategic interventions to ongoing conflicts to other donors actions and the efforts of various groups in the country to obtain or hold on to power—that macro-level indicators are a misleading guide to whether or not USAID projects are in fact having an impact. USAID efforts in such areas as strengthening election commissions, building independent media, or supporting opposition political parties may be successful at the project level but only become of vital importance to changing overall levels of democracy much later, when other factors internal to the country's political processes open opportunities for political change (McFaul 2006). Learning "what works" requires that USAID focus its efforts to gather and analyze data on outcomes at the appropriate level for evaluating specific projects—what is labeled "outcome" measures in Table 2-1.

The committee wants to stress that there are good reasons for employing meso- and macro-level indicators of democracy and working to improve them. They are important tools for strategic assessment of a country's current condition and long-term trajectory regarding democratization. But these indicators are usually not good tools for project evaluation. For the latter purpose, what is needed, as Sarles noted, are

measures that are both policy relevant and plausibly linked to a specific policy intervention sponsored by USAID. The committee discusses these policy-relevant outcome measures and provides examples from our field visits in Chapter 7.

If one concern regarding USAID's evaluation processes is that they may rely too much on meso- and macro-measures to judge program success, the committee also found a related concern regarding USAID's data collection for M&Es: USAID spends by far the bulk of its M&E efforts on data at the "output" level, the first category in Table 2-1.

Current M&E Practices and the Balance Among Types of Evaluations

In the current guidelines for USAID's M&E activities given earlier, only monitoring is presented as "an ongoing, routine effort requiring data gathering, analysis, and reporting on results at periodic intervals." Evaluation, by contrast, is presented as an "occasional" activity to be undertaken "only when needed." The study undertaken for SORA by Bollen et al (2005) that is discussed in Chapter 1 found that most USAID evaluations were process evaluations. These can provide valuable information and insights but, as already discussed, do not help assess whether a project had its intended impact.

Although we cannot claim to have made an exhaustive search, the committee asked repeatedly for examples of impact evaluations for DG projects. The committee learned about very few. One example was a well-designed impact evaluation of a project to support CSOs in Mali (Management Systems International 2000). Here the implementers had persuaded USAID to make use of annual surveys being done in the country, and to use those surveys to measure changes in attitudes toward democracy in three distinct areas: those that received the program, those that were nearby but did not receive the program (to check for spillover effects), and areas that were distant from the sites of USAID activity. The results of this evaluation suggested that USAID programs were not having as much of an impact as the implementers and USAID had hoped to see.

The response within USAID was informative. Some USAID staff members were concerned that a great deal of money had been spent to find little impact; complaints were thus made that the evaluation design had not followed changes made while the program was in progress or was not designed to be sensitive to the specific changes USAID was seeking. On the other hand, there were also questions about whether annual surveys were too frequent or too early to capture the results of investments that were likely to pay off only in the longer term. And the project, by funding hundreds of small CSOs, might have suffered from its own design flaws; some of those who took part in the project suggested that fewer

and larger investments in a select set of CSOs might have had a greater impact. All of these explanations might have been explored further as a way to understand when and how impact evaluations work best. But from the committee's conversations, the primary "lessons" taken away by some personnel at USAID were that such rigorous impact evaluations were not worth the time, effort, and money given what they expected to get from them or did not work.

While certainly only a limited number of projects should be subject to full evaluations, proper impact evaluations cannot be carried out unless "ongoing and routine efforts" to gather appropriate data on policy-relevant outcomes before, during, and after the project are designed into an M&E plan from the inception of the project. Current guidelines for M&E activity tend to hinder making choices between impact and process evaluations and in particular make it very difficult to plan the former. Chapter 7 discusses, based on the committee's field visits to USAID DG missions, the potential for improving, in some cases, USAID M&E activities simply by focusing more efforts on obtaining data at the policy outcome level.

Using Evaluations Wisely: USAID as a Learning Organization

Even if USAID were to complete a series of rigorous evaluations with ideal data and obtained valuable conclusions regarding the effectiveness of its projects, these results would be of negligible value if they were not disseminated through the organization in a way that led to substantial learning and were not used as inputs to planning and implementation of future DG projects. Unfortunately, much of USAID's former learning capacity has been reduced by recent changes in agency practice.

A longstanding problem is that much project evaluation material is simply maintained in mission archives or lost altogether (Clapp-Wincek and Blue 2001). For example, the committee found that when project evaluations involved surveys, while the results might be filed in formal evaluation reports, the underlying raw data were discarded or kept by the survey firm after the evaluation was completed. While many case studies of past projects, as well as many formal evaluations, are supposed to be available to all USAID staff online, not all evaluations were easy to locate. Moreover, simply posting evaluations online does not facilitate discussion, absorption, and use of lessons learned. Without a central evaluation office to identify key findings and organize conferences or meetings of DG officers to discuss those findings, the information is effectively lost.

As mentioned above, CDIE is no longer active. USAID also formerly had conferences of DG officers to discuss not only CDIE-sponsored evaluations but also research and reports on DG assistance undertaken by

NGOs, academics, and other donors. These activities appear to have significantly atrophied. The committee is concerned about the loss of these learning activities. Even the best evaluations will not be used wisely if their lessons are not actively discussed and disseminated in USAID and placed in the context of lessons learned from other sources, including research on DG assistance from outside the agency and the experience of DG officers themselves. The committee discusses the means to help USAID become a more effective learning organization in Chapters 8 and 9.

CONCLUSIONS

This review of current evaluation practices regarding development assistance in general and USAID's DG programs in particular leads the committee to a number of findings:

- The use of impact evaluations to determine the effects of many parts of foreign assistance, including DG, has been historically weak across the development community. Within USAID the evaluations most commonly undertaken for DG programs are process and participatory evaluations; impact evaluations are a comparatively underutilized element in the current mix of M&E activities.
- Some donors and international agencies are beginning to implement more impact evaluations. Nonetheless, considerable concerns and skepticism remain regarding the feasibility and appropriateness of applying impact evaluations to DG projects. These need to be taken seriously and addressed in any effort to introduce them to USAID.
- Current practices regarding measurement and data collection show a tendency to emphasize collection of output measures rather than policy-relevant outcome measures as the core of M&E activities. There is also a tendency, in part because of the lack of good meso-level indicators, to judge the success of DG programs by changes in macro-level measures of a country's overall level of democracy, rather than by achieving outcomes more relevant to a project's plausible impacts.
- Much useful information aside from evaluations, such as survey data and reports, detailed spending breakdowns, and mission director and DG staff reports, remains dispersed and difficult to access.
- USAID has made extensive investments in developing outcome measures across all its program areas; these provide a sound basis for improving measurements of the policy-relevant effects of DG projects.
- Once completed, there are few organizational mechanisms for broad discussion of USAID evaluations among DG officers or for integra-

tion of evaluation findings with the large range of research on democracy and democracy assistance being carried on outside the agency.

- Many of the mechanisms and opportunities for providing organizational learning were carried out under the aegis of the CDIE. The dissolution of this unit, combined with the longer term decline in regular evaluation of projects, means that USAID's capacity for drawing and sharing lessons has disappeared. The DG office's own efforts to provide opportunities for DG officers and implementers to meet and learn from one another and outside experts have also been eliminated.

- Evaluation is a complex process, so that improving the mix of evaluations and their use, and in particular increasing the role of impact evaluations in that mix, will require a combination of changes in USAID practices. Gaining new knowledge from impact evaluations will depend on developing good evaluation designs (a task that requires special skills and expertise), acquiring good baseline data, choosing appropriate measures, and collecting data on valid comparison groups. Determining how to feasibly add these activities to the current mix of M&E activities will require attention to the procedures governing contract bidding, selection, and implementation. The committee's recommendations for how USAID should address these issues are presented in Chapter 9.

Moreover, better evaluations are but one component of an overall design for learning, as making the best use of evaluations requires placing the results of all evaluations in their varied contexts and historical perspectives. This requires regular activities within USAID to absorb and disseminate lessons from case studies, field experience, and research from outside USAID on the broader topics of democracy and social change. The committee's recommendations on these issues are presented in Chapter 8.

These recommendations are intended to improve the value of USAID's overall mix of evaluations, to enrich its strategic assessments, and to enhance its capacity to share and learn from a variety of sources—both internal and from the broader community—about what works and what does not in efforts to support democratic progress.

REFERENCES

- Asia Foundation. 2007. Afghanistan in 2007: A Survey of the Afghan People. Available at: <http://www.asiafoundation.org/pdf/AG-survey07.pdf>. Accessed on February 23, 2008.
- Banerjee, A.V. 2007. *Making Aid Work*. Cambridge, MA: MIT Press.
- Bollen, K., Paxton, P., and Morishima, R. 2005. Assessing International Evaluations: An Example from USAID's Democracy and Governance Programs. *American Journal of Evaluation* 26:189-203.

- CIDA (Canadian International Development Agency). 2007. Results-Based Management in CIDA: An Introductory Guide to the Concepts and Principles. Available at: <http://www.acdi-cida.gc.ca/CIDAWEB/acdicida.nsf/En/EMA-218132656-PPK#1>. Accessed on September 12, 2007.
- Clapp-Wincek, C., and Blue, R. 2001. *Evaluation of Recent USAID Evaluation Experience*. Washington, DC: USAID, Center for Development Information and Evaluation.
- Danish Ministry of Foreign Affairs. 2005. *Peer Assessment of Evaluation in Multilateral Organizations: United Nations Development Programme*, by M. Cole et al. Copenhagen: Ministry of Foreign Affairs of Denmark.
- DfID (Department for International Development). 2004. Public Information Note: Drivers of Change. Available at: <http://www.gsdrc.org/docs/open/DOC59.pdf>. Accessed on September 16, 2007.
- Green, A.T., and Kohl, R.D. 2007. Challenges of Evaluating Democracy Assistance: Perspectives from the Donor Side. *Democratization* 14(1):151-165.
- House of Commons (Canada). 2007. *Advancing Canada's Role in International Support for Democratic Development*. Ottawa: Standing Committee on Foreign Affairs and International Development.
- Jacquet, P. 2006. Evaluations and Aid Effectiveness. In *Rescuing the World Bank: A CGD Working Group Report and Collected Essays*, N. Birdsall, ed. Washington, DC: Center for Global Development.
- Kessler, G. 2007. Where U.S. Aid Goes Is Clearer, But System Might Not Be Better. *Washington Post*, p. A1.
- McFaul, M. 2006. *The 2004 Presidential Elections in Ukraine and the Orange Revolution: The Role of U.S. Assistance*. Washington, DC: USAID, Office for Democracy and Governance.
- McMurtry, V.A. 2005. *Performance Management and Budgeting in the Federal Government: Brief History and Recent Developments*. Washington, DC: Congressional Research Service.
- Management Systems International. 2000. Third Annual Performance Measurement Survey: Data Analysis Report. USAID/Mali Democratic Governance Strategic Objective. Unpublished.
- Millennium Challenge Corporation. 2007. Fiscal Year 2007 Guidance for Compact Eligible Countries, Chapter 29, Guidelines for Monitoring and Evaluation Plans, p. 19. Available at: <http://www.mcc.gov/countrytools/compact/fy07guidance/english/29-guidelinesformande.pdf>. Accessed on September 12, 2007.
- OECD (Organization for Economic Cooperation and Development). 2005. Lessons Learned on the Use of Power and Drivers of Change Analyses in Development Operation. Review commissioned by the OECD DAC Network on Governance, Executive Summary. Available at: <http://www.gsdrc.org/docs/open/DOC82.pdf>. Accessed on September 12, 2007.
- Sarles, M. 2007. Evaluating the Impact and Effectiveness of USAID's Democracy and Governance Programmes, in *Evaluating Democracy Support: Methods and Experiences*, P. Burnell, ed. Stockholm: International Institute for Democracy and Electoral Assistance and Swedish International Development Cooperation Agency.
- Savedoff, W.D., Levine, R., and Birdsall, N. 2006. *When Will We Ever Learn? Improving Lives Through Impact Evaluation*. Washington, DC: Center for Global Development.
- Schmid, A. 2007. Measuring Development. Available at: <http://www.gtz.de/de/dokumente/ELR-en-30-31.pdf>. Accessed on September 12, 2007.
- Shadish, W.R., Cook, T.D., and Campbell, D.T. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, 2nd ed. Boston: Houghton Mifflin.
- USAID ADS. 2007. Available at: <http://www.usaid.gov/policy/ads/200/>. Accessed on August 2, 2007.
- USAID (U.S. Agency for International Development). 1997. *The Role of Evaluation in USAID. TIPS 11*. Washington, DC: USAID.

- USAID (U.S. Agency for International Development). 1998. *Handbook of Democracy and Governance Program Indicators*. Washington, DC: Center for Democracy and Governance. USAID. Available at: http://www.usaid.gov/our_work/democracy_and_governance/publications/pdfs/pnacc390.pdf. Accessed on August 1, 2007.
- USAID (U.S. Agency for International Development). 2000. *Conducting a DG Assessment: A Framework for Strategy Development*. Available at: http://www.usaid.gov/our_work/democracy_and_governance. Accessed on August 26, 2007.
- USAID (U.S. Agency for International Development). 2006. U.S. Foreign Assistance Reform. Available at: http://www.usaid.gov/about_usaid/dfa/. Accessed on August 2, 2007.
- USAID (U.S. Agency for International Development). 2007. Decentralization and Democratic Local Governance (DDLG) Handbook. Draft.
- U.S. Department of State. 2006. U.S. Foreign Assistance Performance Indicators for Use in Developing FY2007 Operational Plans, Annex 3: Governing Justly and Democratically: Indicators and Definitions. Available at: <http://www.state.gov/ff/releases/factsheets2007/78450.htm>. Accessed on August 25, 2007.
- Wholey, J.S., Hatry, H.P., and Newcomer, K.E., eds. 2004. *Handbook of Practical Program Evaluation*, 2nd ed. San Francisco: Jossey-Bass.
- World Bank. 2004. *Monitoring & Evaluation: Some Tools, Methods, and Approaches*. Washington, DC: World Bank.
- de Zeeuw, J., and Kumar, K. 2006. *Promoting Democracy in Postconflict Societies*. Boulder: Lynne Rienner.

Measuring Democracy¹

INTRODUCTION

One of the U.S. Agency for International Development's (USAID) charges to the National Research Council committee was to develop an operational definition of democracy and governance (DG) that disaggregates the concept into clearly defined and measurable components. The committee sincerely wishes that it could provide such a definition, based on current research into the measurement of democratic behavior and governance. However, in the current state of research, only the beginnings of such a definition can be provided. As detailed below, there is as much disagreement among scholars and practitioners about how to measure democracy, or how to disaggregate it into components, as on any other aspect of democracy research. The result is that there exist a welter of competing definitions and breakdowns of "democracy," marketed by rivals, each claiming to be a superior method of measurement, and each the subject of sharp and sometimes scathing criticism.

The committee believes that democracy is an inherently multidimensional concept, and that broad consensus on those dimensions and how

¹Helpful comments on this chapter were received from Macartan Humphreys, Fabrice Lehoucq, and Jim Mahoney. The committee is especially grateful to those who attended a special meeting on democracy indicators held at Boston University in January 2007: David Black, Michael Coppedge, Andrew Green, Rita Guenther, Jonathan Hartlyn, Jo Husbands, Gerardo Munck, Margaret Sarles, Fred Schaffer, Richard Snyder, Paul Stern, and Nicolas van de Walle. See Appendix C for further information.

to aggregate them may never be achieved. Thus, if USAID is seeking an operational measure of democracy to track changes in countries over time and where it is engaged, a more practical approach would be to disaggregate the various components of democracy and track changes in democratization by looking at changes in those components.

Yet even for the varied components of democracy, there are no available measures that are widely accepted and have demonstrated the validity, accuracy, and sensitivity that would make them useful for USAID in tracking modest changes in democratic conditions in specific countries. The development of a widely recognized disaggregated definition of democracy, with clearly defined and objectively measurable components, would be the result of a considerable research project that is yet to be done.

This chapter provides an analysis of existing measures of democracy and points the way toward developing a disaggregated measure of the type requested by USAID. The committee finds that most existing measures of democracy are adequate, and in fair agreement, at the level of crude determination of whether a country is solidly democratic, autocratic, or somewhere in between. However, the committee also finds that all existing measures are severely inadequate at tracking small movements or small differences in levels of democracy between countries or in a single country over time. Moreover, the committee finds that existing measures disaggregate democracy in very different ways and that their measures of various components of democracy do not provide transparent, objective, independent, or reliable indicators of change in those components over time.

While recognizing that it may seem self-serving for an academic committee to recommend “more research,” it is the committee’s belief—after surveying the academic literature and convening a workshop of experts in democracy measures to discuss the issue—that if USAID wishes a measure of democracy that it can use to gauge the impact of its programs and track the progress of countries in which it is active, it faces a stark choice: either rely on the current flawed measures of democracy or help support the development of a research project on democracy indicators that—it is hoped—will eventually produce a set of indicators with the broadly accepted integrity of today’s national accounts indicators for economic development.

To provide just a few examples to preview the discussion below, USAID manages its DG programs with an eye toward four broad areas: rule of law, elections, civil society, and good governance. Yet consider the two most widely used indicators of democracy: the Polity autocracy/democracy scale and the Freedom House scales of civil liberties and political rights. The former breaks down its measures of democracy into three components: executive recruitment, executive constraints, and political

competition, measured by six underlying variables. While some of these could be combined to provide indicators of elections, civil society, and aspects of rule of law, Polity does not address “good governance.” Moreover, the validity of the various components and underlying variables in Polity is so greatly debated that there is no reason to believe that a measure of rule of law based on the Polity components would be accepted. Freedom House rates nations on two scales: civil liberties (which conflates rule of law, civil society, and aspects of good governance) and political rights (which conflates rule of law, elections, and aspects of good governance). Even if these scales were based on objective and transparent measurements (and they are not), there would be no way to extract from them information on components relevant to USAID’s DG policy areas.

Fortunately, while more sensitive and accurate measures to track sectoral movements toward or away from democracy are vital to improving USAID’s policy planning and targeting of DG programs, USAID can still gain knowledge on the impacts of its programs by focusing on changes in outcome indicators at a level relevant to those projects (for which methodologies are examined in Chapters 5 through 7). That is, USAID should seek to determine whether its projects lead to more independent and effective behavior by judges and legislators, broader electoral participation and understanding by citizens, more competitive and fair election practices, fewer corrupt officials, and other concrete changes. The issue of how much those changes contribute to overall trajectories of democracy or democratic consolidation is one that can only be solved by future experience and study and the development of better disaggregated measures for tracking democracy at the sectoral level.

The committee thus agrees that USAID is correct in focusing its interest in measurement on developing a measure of democracy that is disaggregated into discrete and measurable components. This chapter will analyze existing approaches to measuring democracy, identifying why they are flawed, and point the way toward what the committee believes will be a more useful approach to developing disaggregated sectoral or meso-level measures (Table 2-1).

PROBLEMS WITH EXTANT INDICATORS

A consensus is growing within the scholarly community that existing indicators of democracy are problematic.² These problems may be grouped into five categories: (1) problems of definition, (2) sensitivity issues, (3) measurement errors and data coverage, (4) aggregation prob-

²See Bollen (1993), Beetham (1994), Gleditsch and Ward (1997), Bollen and Paxton (2000), Foweraker and Krznaric (2000), McHenry (2000), Munck and Verkuilen (2002), Treier and Jackman (2003), Berg-Schlosser (2004 a, b), Acuna-Alfaro (2005), and Vermillion (2006).

lems, and (5) lack of convergent validity. What follows is a brief, sometimes rather technical, review of these problems and their repercussions. Definitions of key terms are provided in the text or in the Glossary at the end of the report.

The focus of the discussion is on several leading democracy indicators: (1) Freedom House; (2) Polity; (3) ACLP (“ACLP” stands for the names of the creators—Alvarez, Cheibub, Limongi, and Przeworski; Alvarez et al 1996; recently expanded by Boix and Rosato 2001); and (4) the Economist Intelligence Unit (EIU). Freedom House provides two indices: “Political Rights” and “Civil Liberties” (sometimes employed in tandem, sometimes singly). Both are seven-point scales extending back to 1972 and cover most sovereign and semisovereign nations.³ Polity also provides two aggregate indices: “Democracy” and “Autocracy.” Both are 10-point scales and are usually used in tandem (by subtracting one from the other), which provides the 21-point (–10 to 10) Polity2 variable. Coverage extends back to 1800 for sovereign countries with populations greater than 500,000.⁴ ACLP codes countries dichotomously (autocracy/democracy) and includes most sovereign countries from 1950 to 1990. The expanded dataset provided by Boix and Rosato (2001) stretches back to 1800.⁵ The EIU has recently developed a highly disaggregated index of democracy with 5 core dimensions and 60 subcomponents, which are combined into a single index of democracy (Kekic 2007). Coverage extends to 167 sovereign or semisovereign nations but only in 2006.

Glancing reference will be made to other indicators in an increasingly crowded field,⁶ and many of the points made in the following discussion apply quite broadly. However, it is important to bear in mind that each indicator has its own particular strengths and weaknesses. The following brief survey does not purport to provide a comprehensive review.⁷

³See www.freedomhouse.org.

⁴Both are drawn from the most recent iteration of this project, known as Polity IV. See www.cidcm.umd.edu/inscr/polity.

⁵Jose Cheibub and Jennifer Ghandi are currently engaged in updating the ACLP dataset, but results are not yet available.

⁶See Bollen (1980), Coppedge and Reinicke (1990), Arat (1991), Hadenius (1992), Vanhanen (2000), Altman and Pérez-Liñán (2002), Gasiorowski (1996; updated by Reich 2002 [also known as “Political Regime Change—PRC dataset”]), and Moon et al (2006).

⁷The most detailed and comprehensive recent reviews are Hadenius and Teorell (2005) and Munck and Verkuilen (2002). See also Bollen (1993), Beetham (1994), Gleditsch and Ward (1997), Bollen and Paxton (2000), Elkins (2000), Foweraker and Krznaric (2000), McHenry (2000), Casper and Tufis (2003), Treier and Jackman (2003), Berg-Schlosser (2004a, b), Acuna-Alfaro (2005), and Bowman et al (2005).

Definition

There are many ways to define democracy, and each naturally generates a somewhat different approach to measurement (Munck and Verkuijlen 2002). Some definitions are extremely “thin,” focusing mainly on the presence of electoral competition for national office. The ACLP index exemplifies this approach: Countries that have changed national leadership through multiparty elections are democracies; other countries are not. Other definitions are rather “thick,” encompassing a wide range of social, cultural, and legal characteristics well beyond elections. For example, the Freedom House Political Rights Index includes the following questions pertaining to corruption:

Has the government implemented effective anticorruption laws or programs to prevent, detect, and punish corruption among public officials, including conflict of interest? Is the government free from excessive bureaucratic regulations, registration requirements, or other controls that increase opportunities for corruption? Are there independent and effective auditing and investigative bodies that function without impediment or political pressure or influence? Are allegations of corruption by government officials thoroughly investigated and prosecuted without prejudice, particularly against political opponents? Are allegations of corruption given wide and extensive airing in the media? Do whistle-blowers, anticorruption activists, investigators, and journalists enjoy legal protections that make them feel secure about reporting cases of bribery and corruption? What was the latest Transparency International Corruption Perceptions Index score for this country? (Freedom House 2007)

It may be questioned whether these aspects of governance, important though they may be, are integral components of *democracy*.

More generally, many scholars treat good governance as a likely *result* of democracy; yet many donors (including USAID) treat good governance as an essential *component* of democracy. Similar complaints might be registered about other concepts and scales of democracy; some are so “thick” as to include diverse elements of accountability, even distributional equity and economic growth.

For example, some definitions treat the United States as a democracy from the passage of its Constitution and first national election in 1789. Yet since George Washington ran uncontested in both 1789 and 1792, even ACLP would not treat the United States as democratic until the appearance of contested multiparty elections in 1796. If slavery is considered a contravention of democracy, the United States could not be considered a democracy until its abolition throughout its territory in 1865. If women’s right to vote is also considered essential to the definition of democracy, the United States does not qualify until 1920. And if the disenfranchisement of African Americans in southern states is considered a block to democ-

racy, the United States does not become a full democracy until passage of the Civil Rights Act in 1965.

In short, only a “thin” definition of democracy would classify the United States as “fully democratic” from the early nineteenth century. Yet most donor agencies are reluctant to adopt such thin measures as a guide to current democracy assessments, questioning whether “thin” indices of democracy capture all the critical features of this complex concept. The problem of definition is critical but very difficult to resolve.

Sensitivity

A related issue is that many of the leading democracy indicators are not sensitive to important gradations in the quality of democracy across countries or through time. At the extreme, dichotomous measures such as ACLP reduce democracy to a dummy variable: A country either is or is not a democracy, with no intermediate stages permitted. While useful for certain purposes, one may wonder whether this captures the complexity of such a variegated concept (Elkins 2000). At best it captures one or two dimensions of democracy (those employed as categorizing principles), while the rest are necessarily ignored.

Most democracy indicators allow for a more elongated scale. As noted above, Freedom House scores democracy on a seven-point index (14 points if the Political Rights and Civil Liberties indices are combined). Polity provides a total of 21 points if the Democracy and Autocracy scales are merged into the Polity2 variable, which gives the impression of considerable sensitivity. In practice, however, country scores stack up at a few places (notably, 7 for autocracies and +10 for full democracies, the highest possible score), suggesting that the scale is not as sensitive as it purports to be. The EIU index is by far the most sensitive and does not appear to be arbitrarily “bunched.”⁸

Note that all extant indicators are bounded to some degree and therefore constrained. This means that there is no way to distinguish the quality of democracy among countries that have perfect negative or positive scores. This is fine as long as there really is no difference in the quality of democracy among these countries. Yet the latter assumption is highly questionable. Consider that in 2004, Freedom House assigned the highest score (1) on its Political Rights Index to the following 58 countries: Andorra, Australia, Austria, Bahamas, Barbados, Belgium, Belize, Bulgaria, Canada, Cape Verde, Chile, Costa Rica, Cyprus (Greek),

⁸Questions can also be raised about whether these indices are properly regarded as interval scales (Treier and Jackman 2003). The committee does not envision an easy solution to this problem.

Czech Republic, Denmark, Dominica, Estonia, Finland, France, Germany, Greece, Grenada, Hungary, Iceland, Ireland, Israel, Italy, Japan, Kiribati, Latvia, Liechtenstein, Luxembourg, Malta, Marshall Islands, Mauritius, Micronesia, Nauru, Netherlands, New Zealand, Norway, Palau, Panama, Poland, Portugal, San Marino, Slovakia, Slovenia, South Africa, South Korea, Spain, St. Kitts and Nevis, St. Lucia, Suriname, Sweden, Switzerland, Tuvalu, United Kingdom, United States, and Uruguay.⁹ Are we really willing to believe that there are no substantial differences in the quality of democracy among these diverse polities?

Measurement Errors and Data Coverage

Democracy indicators often suffer from measurement errors and/or missing data.¹⁰ Some (e.g., Freedom House) are based largely on expert judgments, judgments that may or may not reflect facts on the ground.¹¹ Some (e.g., Freedom House in the 1970s and 1980s) rely heavily on secondary accounts from a few newspapers such as the *New York Times*. These accounts may or may not be trustworthy and almost assuredly do not provide comprehensive coverage of the world. Moreover, newspaper accounts suffer from extreme selection bias, depending almost entirely on the location of the newspaper's reporters. Thus, if the *New York Times* has a reporter in Mexico but none in Central America, coverage of the latter is going to be much spottier than the former. In an attempt to improve coverage and sophistication, some indices (e.g., EIU) impute a large quantity of missing data. This is a dubious procedure wherever data coverage is limited, as it seems to be for many of the EIU variables. Note that many of the EIU variables rely on polling data, which are available on a highly irregular basis for 100 or so nation states.

The quality of many of the surveys on which the EIU draws has not been clearly established. This means that data for these questions must be estimated by country experts for all other cases, estimated to be about half of the sample. (The procedures employed for this estimation are not known.)

Wherever human judgments are required for coding, one must be

⁹The precise period in question stretches from December 1, 2003, to November 30, 2004; obtained from <http://www.freedomhouse.org/template.cfm?page=15&year=2006> (accessed on September 21, 2006).

¹⁰For general treatments of the problem of conceptualization and measurement, see Adcock and Collier (2001).

¹¹With respect to the general problem of expert judgments, see Tetlock (2005), who found that expert opinions tended to reflect more the consensus of the expert community than an objective "truth," inasmuch as his surveys of experts produced answers that were often, in retrospect, no more accurate than a coin toss.

concerned about the basis of the respondent's decisions. In particular, one wonders whether coding decisions about particular topics (e.g., press freedom) may reflect an overall sense to outside experts of how democratic country A is, rather than an independent evaluation of the question at hand. The committee also worries about the problem of endogeneity of the evaluations, that is, with experts looking more at what other experts and indicators are doing rather than making their own independent evaluation of the country. The intercoder "reliability" may be little more than an artifact of experts accepting other experts' judgments. In this respect, "disaggregated" indicators are often considerably less disaggregated than they appear. Note that it is the ambiguity of the questionnaires underlying these surveys that fosters this sort of premature aggregation.

The committee undertook a limited statistical examination of the Freedom House scores for 2007 on their key components—for political rights this included electoral process, pluralism and participation, and functioning of government; for civil liberties these were freedom of expression, association and organizational rights, rule of law, and personal autonomy and individual rights (see Appendix C). Across all countries, two-way correlations among the seven components were never less than 0.86 and in several cases were 0.95 or greater. This high correlation could imply that democracy is indeed a far "smoother" condition than the "lumpy" view expressed in this study. That is, the high correlation among the items suggests that picking any one is just about as good as picking any other. Yet the committee doubts the independence of the judgments on each of the components of the scale.

The EIU democracy scale also is divided into components: civil rights, elections, functioning of government, participation, and culture. Taking the Freedom House and EIU components together, a factor analysis reveals that a single factor loading explains 83 percent of the variance across all 12 components, and the two principal factors explain 90 percent of the variance (Coppedge 2007). This, by itself, is not problematic; it could be that good/bad things go together; that is, countries that are democratic on one dimension are also democratic on another. However, it raises concern about the actual independence of the various components in these indices. It could be, in other words, that respondents (either experts or citizens) who are asked about different dimensions of a polity are, in fact, simply reflecting their overall sense of a country's democratic culture. It also suggests that the various independent components in fact contain no more useful information than the principal one or two factors.

Adding to worries about measurement error is the general absence of intercoder reliability tests as part of the coding procedure. Freedom House does not conduct such tests (or at least does not make them public). Polity does so, but it requires a good deal of hands-on training before coders reach an acceptable level of coding accuracy. This suggests that other cod-

ers would not reach the same decisions simply by reading Polity's coding manual or that artificial uniformity is imposed. And this, in turn, points to a potential problem of conceptual validity: Key concepts are not well matched to the empirical data.

Aggregation

Since democracy is a multifaceted concept, all composite indicators must wrestle with the aggregation problem—how to weight the components of an index and which components to include. For aggregation to be successful, the rules must be clear, operational, and consistent with common notions of what democracy is; that is, the resulting concept must be valid. It goes almost without saying that different solutions to the aggregation problem lead to quite different results (Munck and Verkuilen 2002; for a possible exception to this dictum, see Coppedge and Reinicke 1990).

Although most indicators have fairly explicit aggregation rules, they are often difficult to comprehend, and consequently to apply. They may also include “wild card” elements, allowing the coder free rein to assign a final score that accords with his or her overall impression of a country (e.g., Freedom House). In some cases (e.g., Polity), components are listed separately, which helps clarify the final score a country receives. However, in Polity's case the components of the index are themselves highly aggregated, so the overall clarity of the indicator is not improved.

Even when aggregation rules are clear and unambiguous, because they bundle a host of diverse dimensions into a single score, it is often unclear which of the dimensions is driving a country's score in a particular year. It is often difficult to articulate what an index value of “4” means within the context of any single indicator.

Moreover, even if an aggregation rule is explicit and operational, it is never above challenge. The Polity index, in Munck and Verkuilen's estimation, “is based on an explicit but nonetheless quite convoluted aggregation rule” (2002:26). Indeed, a large number of possible aggregation rules fit, more or less, with everyday concepts of democracy and thus meet the minimum requirements of conceptual validity. For this reason the committee regards the aggregation problem as the only problem that is unsolvable *in principle*. There will always be disagreement over how to aggregate the various components of “Big D democracy” (i.e., the one central concept that is assumed to summarize a country's regime status).

Convergent Validity

Given the above, it is no surprise that there is significant disagreement among scholars over how to assign scores for particular countries on

the leading democracy indices. Granted, intercorrelations among various democracy indicators are moderately high, suggesting some basic agreement over what constitutes a democratic state. As shown in the analysis undertaken for the committee that is summarized in Appendix C, the Polity2 variable (combining Democracy and Autocracy) drawn from the Polity dataset and the Freedom House Political Rights Index are correlated at .88 (Pearson's r). Yet when countries with perfect democracy scores (e.g., the United Kingdom and the United States) are excluded from the samples, this intercorrelation drops to .78. And when countries with scores of 1 and 2 on the Freedom House Political Rights scale (the two top scores) are eliminated, the correlation experiences a further drop—to .63, implying that two-thirds of the variance in one scale is unrelated to changes in the other scale for countries outside the upper tier of democracies.

The committee similarly finds that correlations between the Freedom House and EIU scores are low when the highest-scoring countries are set aside. For a substantial number of countries—Ghana, Niger, Guinea-Bissau, the Central African Republic, Chad, Russia, Cambodia, Haiti, Cuba, and India—the Freedom House and EIU scores differ so widely that they would be considered democratic by one scale but undemocratic by the other. Indeed, country specialists often take issue with the scoring of countries they are familiar with (e.g., Bowman et al 2005; for more extensive cross-country tests, see Hadenius and Teorell 2005).

Since tracking progress in democracy assistance often depends on accurately measuring modest improvements in democracy, it is particularly distressing that the convergence between different scales is so low in this regard. While the upper “tails” of the distributions on the major indicators (the fully democratic regimes) are highly correlated, the democracy scores for countries in the upper middle to the bottom ranges are not. The analysis commissioned by the committee (see Appendix C) found that the average correlation between the annual Freedom House and Polity scores for autocratic countries (those with Polity scores less than -6) during 1972-2002 was only .274. Among the partially free countries of the former Soviet Union, the correlation between annual Freedom House and Polity scores for the years 1991-2002 was .295; for the partially free countries in the Middle East, it was 0.40. In many cases the correlations for specific countries were *negative*, meaning that the two scales gave opposite measures of whether democracy levels were improving or not. This is a serious problem for USAID and other donors, since they are generally most concerned with identifying the level of democracy, and degrees of improvement, *precisely* for those countries lying in the middle and bottom of the distribution—countries that are mainly undemocratic or imperfectly democratic—rather than for countries already at the upper end of the democracy scale.

If there is little agreement on the quality and direction of democracy in countries that lie in between the extremes, it must be concluded that there is relatively little convergent validity across the most widely used democracy indicators. That is, whatever their intent, they are not in fact capturing the same concept.

By way of conclusion to this very short review of extant indicators, the committee quotes from another recent review by Jim Vermillion, current executive vice president of the International Foundation for Election Systems:

Initial work in the measurement of democracy has provided some excellent insights into specific measures and has helped enlighten our view of where underlying concepts related to democracy stand. However, we are far from coming up with a uniform, theoretically cohesive definition of the construct of democracy and its evolution that lends itself easily to statistical estimation/manipulation and meaningful hypothesis testing. (Vermillion 2006:30)

The need for a new approach to this ongoing, and very troublesome, problem of conceptualization and measurement is apparent.

Average Versus Country-Specific Results

It is reasonable to ask, if the existing indicators of democracy have so many problems, how can the committee have any confidence in the findings mentioned in Chapter 1, such as that the number of democracies in the world is rising and that USAID DG assistance has, on average, made a significant positive difference in democracy levels? For that matter, how is it possible for scholars to have undertaken more than two decades of quantitative research on democracy and democratization, correlating various causal factors with shifts in these democracy indicators, with any belief in the validity of their research?

The answers to this question lie in the very different purposes that democracy indicators must serve for scholarly analysis of average or overall global trends, as against the purposes they must serve to support policy analysis of trends *in specific countries*. For the former purpose it is acceptable for democracy data to have substantial errors regarding levels of democracy in particular states, as long as the errors are not systematically biased. That is, even a democracy scale that makes substantial errors will be useful for looking at average trends as long as its score for any given country is equally likely to be "too high" or "too low." Such a scale will state the level of democracy as too high in about half the world's countries and too low in the other half, but the average level of global democracy overall will be fairly correct, and scholars can use statistical methods to "separate out" the random errors from the overall trends.

Statistical analyses of democracy that use extant indicators such as Polity or Freedom House are looking for the *overall or average* effects of various factors—such as economic growth, democracy assistance, or regime types—on democracy. Thus the Finkel et al studies (2007, 2008) described above, which demonstrate a positive impact of various forms of democracy assistance on *average* levels of democracy while statistically controlling for a host of background, trend, and other causal variables, also controlled for measurement errors in the democracy indices that were assumed to be evenly distributed across countries. What their results tell us is something like the following: In any four-year period, if three countries are examined in which USAID invested an average of \$10 million per country per year in DG assistance, those countries' Freedom House scores will show an overall increase of three points (an average increase of one point per country) at the end of those four years relative to what would have been expected in the absence of USAID DG assistance.¹² Let us accept this finding as the best available estimate of the truth (and this study has been subjected to careful peer criticism and its results stand up well)—*on average*, DG programs do achieve positive results.

Yet such measures are not helpful, indeed can even be misleading, if used to evaluate the effects of DG programming in particular countries. For example, say that USAID spends \$10 million on various DG programs in each of three countries. Say also that a valid and accurate democracy scale (assuming we are able to set aside the effects of any other factors on levels of democracy) would show that such programs led country 1 to increase by two points on this democracy scale and country 2 to increase by one point, while country 3 saw no change. USAID assistance programs thus achieved substantial success in one case, modest success in another, and no effect in the last.

However, the flawed indicator we have instead records that country 1 increased by three points and country 2 *decreased* by one point, while country 3 increased by one point. *On average*, this is exactly the same result—overall scores in these countries increased by a total of three points (or an average of one point per country) for these countries over four years. Yet if USAID relies on this flawed indicator to estimate the impact of its efforts *in specific countries*, it will be considerably off. It will greatly overestimate the success of its programs in countries 1 and 3 and wrongly conclude that its programs were associated with a *decline* in democracy in country 2—all of this just because of random errors in the way that current democracy indicators track small movements or middle-range levels of democracy in particular countries. If USAID were

¹²Finkel et al (2007, 2008) found essentially the same results with Polity scores as Freedom House scores, so this discussion holds for both indicators.

then to ramp up and spread the program in country 1, thinking it an overwhelming (rather than modest) success, and also spread the programs in country 3 that “seemed” to produce a success, while halting the programs that apparently failed to stem democracy decline in country 2, it could be making severe mistakes. Thus the errors found in current widely used democracy indicators, while still allowing them to serve well enough for purposes of scholarly research on average effects of various factors on democracy or for charting overall democracy trends, do *not* serve USAID at all well for the policy purposes of determining the effects of specific programs in particular countries.¹³

For this reason the rest of this chapter lays out an approach that the committee believes will be more fruitful for developing useful indicators of democratic change. Also for this reason, throughout this report methods are stressed for helping USAID determine the effects of its programs using more concrete indicators of the immediate policy outcomes of those programs, rather than macrolevel indicators of national levels of democracy.

A DISAGGREGATED APPROACH TO MEASUREMENT AT THE COUNTRY LEVEL

Given the multiple difficulties encountered by Freedom House, Polity, ACLP, EIU, and other extant indicators of democracy, one might reasonably conclude that the stated task simply cannot be accomplished. That is, one cannot assign a single point score to a particular country at a particular point in time, expecting that this score will accurately capture all the nuances of democracy and be empirically valid through time and across space. The goal of precise numerical comparison is impossible.

While this conclusion may seem compelling, at least initially, one should also consider the costs of *not* comparing in a systematic fashion. Without some way of analyzing the quality of democracy through time and across countries, there is no way to mark progress or regress on this vital factor, to explain it, or to affect its future course. To gain knowledge of the world, and hence to make effective policy interventions, comparisons must be made. And to compare with precision numerical scores must be assigned to countries according to the quality of democracy they sup-

¹³As discussed in Chapter 4, when scholars undertake case studies of democratization in a particular country, they generally do not bother with indicators such as Polity or Freedom House to describe trends in that country, but instead focus on institutional or behavioral changes that they document in detail and seek the causes or consequences of those observed changes.

posedly possess.¹⁴ How, then—given the shortcomings of extant democracy indices—might this difficult task be handled more effectively?

The committee proposes that the key to developing a more accurate and useful empirical approach to democracy—as to other large and unwieldy subjects (e.g., “governance”)—is to be found in greater disaggregation (Coppedge, forthcoming). Rather than focusing on how, precisely, to define democracy and attempting to arrive at a summary score (à la Freedom House or Polity), the committee proposes to focus on developing the most transparent, independent, and valid measures for the underlying dimensions of this concept. The key point is that this approach to data gathering takes place at a much lower level of abstraction than Big D democracy.

Previous Efforts at Disaggregation

The idea of disaggregating measures of democracy and governance is of course not entirely new. As mentioned, the Polity IV dataset includes six component variables, each measured separately. Other precedents include the *Handbook of Democracy and Governance Program Indicators* (USAID 1998), the *Bertelsmann Transformation Index* (Bertelsmann Foundation 2003), the *Database of Political Institutions* (Beck et al 2000), the EIU index (Kekic 2007), and the World Bank governance indicators (Kaufmann et al 2006).

In some areas—for example, free press (Freedom House 2006) or elections (Munck 2006)—disaggregated topics have been successfully measured on a global scale. In these and other instances, the committee suggests building on, or simply incorporating, previous efforts. However, the usual approach to disaggregation is flawed, either because the resulting indicators are still highly abstract and hence difficult to operationalize (e.g., Polity IV) and/or because the underlying components, while conceptually distinct, are gathered in such a way as to compromise their independence.

Consider the six World Bank governance indicators—government effectiveness, voice and accountability, control of corruption, rule of law, regulatory burden, and political instability—which involve very similar underlying components (Landman 2003, Kurtz and Schrank 2007, Thomas 2007). Issues of corruption, for example, figure in several of the six dimensions. It seems likely that overall perceptions on the part of

¹⁴To some the assignment of a point score may seem a prime example of misplaced precision. Yet the lack of precision inherent in such cross-country comparisons can be handled by including an estimate of uncertainty along with the point estimate so that users of the data will not be misled.

survey respondents (whether expert or civilian) as to “how country A is doing” color many of the survey responses on which these indicators depend, insofar as survey questions tend to be quite broad. This sort of disaggregation does not achieve the intended purpose. Indeed, it is often argued that the six Kaufmann variables are best regarded as measures of the same thing and therefore are often combined in empirical analyses.

A similar problem besets other efforts at disaggregation, such as the recently released Freedom House measures of civil liberties and political rights, which are broken down into seven components: electoral process, political pluralism and participation, functioning of government, freedom of expression and belief, associational and organizational rights, rule of law, personal autonomy, and individual rights (Freedom House 2007). Again, the extremely high correlations among these components ($>.87$ on all paired comparisons; see Appendix C), along with the vagueness of the questions and coding procedures, prompts us to wonder whether these are truly *independent* measures of democracy, or simply different ways of accessing a country’s overall gestalt.

The EIU index does a slightly better job of disaggregating its component variables, which are reported for five dimensions: electoral process and pluralism, civil liberties, the functioning of government, political participation, and political culture. Correlations are still quite high but not outrageously so. Moreover, the specificity of the questions makes the claim of independence among these five variables plausible. Unfortunately, the committee was not able to get access to the data for the 60 specific questions that compose the five dimensions. It is quite possible that these underlying data are regarded by EIU as proprietary. If so, the index will have much less utility for policy and especially scholarly purposes.

Meanings of Democracy

We turn now to the vexing problem of definition, to which we have already alluded. Democracy means rule by the people, and this core attribute has remained relatively constant since the term was invented by the Greeks. Yet the notion of popular sovereignty is exceedingly vague. Thus, it may be necessary to adopt a more specific definition if the term is to have any practical utility. Unfortunately, in articulating an operational definition of democracy, considerable disagreement is encountered both within and outside the academic community. These disagreements are partly the product of cross-cultural differences (Schaffer 1998). More fundamentally, they are a product of the multiple uses that have developed over many centuries (Dunn 2006).

For current purposes the committee is primarily concerned with the concept as it might be applied to populous communities, that is, to nation-

states, regions, and large municipalities. In this context the term is nowadays frequently identified with political contestation (also often called competition), as secured through an electoral process by which leaders are selected. Where effective competition exists, democracy is also said to exist (Schumpeter 1942, Alvarez et al 1996). For many writers, competition is the *sine qua non* of democracy. This may be regarded as a minimalist (or “thin”) definition of the concept.

Although there is general consensus about the importance of political competition, many other attributes have also been understood as defining features of democracy. These include liberty/freedom, accountability, responsiveness, deliberation, participation, political equality, and social equality. Each of these attributes may in turn be broken down into lower-level components, so the field of potential attributes is indeed quite vast. Adding these attributes to the minimal definition—political competition—various maximalist, or ideal-type, definitions of the concept can be constructed. Arguably, a true, complete, or full democracy should possess all of the foregoing definitional attributes, and each should be fully developed.

Unfortunately, the committee sees no way of resolving the choice between minimal and maximal definitions of democracy. The first seems too small; it excludes too much. But the latter is clearly too large and unwieldy to be serviceable; it is, indeed, indistinguishable from good governance. Moreover, the many possible resolutions of this dilemma that lie in between minimal and maximal definitions cannot avoid the problem of arbitrariness: Why should some elements of democracy (as that concept is commonly employed) be included, while others are excluded? As a general rule, stipulated definitions tend to be poorly bounded, imprecise, or arbitrary (i.e., they violate ordinary usages of the concept and therefore do not “make sense”). The committee realizes that definitions must often be stipulated. But if the resulting indicators are not perceived as legitimate by policymakers and citizens on a global level, they are unlikely to perform the work that USAID and others expect of it. An illegitimate index, particularly one that is considered arbitrary and involves excessive judgment on the part of coders, is easy to dismiss.

Thus, although one of the original tasks given to this committee by USAID was to develop an “initial operational definition of democracy and governance,” as discussed above, the committee has concluded, after extensive consultation among committee members and with leading authorities on democracy, that it is not possible for it to do so. The challenges facing any particular committee of scholars in producing a definition that would command wide assent, as outlined above, are simply too great.

Thirteen Dimensions of Democracy

The committee's proposed solution is to suggest, as a starting point for further study, a disaggregation of the concept of democracy down to a level where greater consensus over matters of definition, along with greater precision of measurement, may be obtained. In this way the committee hopes to sidestep the eternally vexing question of what "democracy" means.

Having considered the matter at some length and having consulted with distinguished experts on the subject, the committee resolved that there are at least 13 dimensions of democracy that are independently assessable (i.e., they do not reduce to some overall conception of "how country A is doing"):

1. **National Sovereignty:** Is the nation sovereign?
2. **Civil Liberty:** Do citizens enjoy civil liberty in matters pertaining to politics?
3. **Popular Sovereignty:** Are elected officials sovereign relative to nonelected elites?
4. **Transparency:** How transparent is the political system?
5. **Judicial Independence:** How independent and empowered is the judiciary?
6. **Checks on the Executive:** Are there effective checks on the executive?
7. **Election Participation:** Is electoral participation unconstrained and extensive?
8. **Election Administration:** Is the administration of elections fair?
9. **Election Results:** Are the results of an election accepted by the citizenry to indicate that a democratic process has occurred?
10. **Leadership Turnover:** Is there regular turnover in the top political leadership?
11. **Civil Society:** Is civil society dynamic, independent, and politically active?
12. **Political Parties:** Are political parties well institutionalized?
13. **Subnational Democracy:** How decentralized is political power and how democratic is politics at subnational levels?

The committee realizes that most of these dimensions are continuous (matters of degree), rather than dichotomous (either/or). Even so, it seems reasonable to refer to them—loosely—as potential *necessary conditions* of a fully democratic polity.

Further details regarding the 13 components of the index, along with some initial suggestions for how to measure them, are discussed in

Appendix C. Here, the reader's attention is called to the following general points:

First, the criteria applying to different dimensions sometimes conflict with one another. For example, strong civil society organizations representing one social group may pressure government to restrict other citizens' civil liberties (Levi 1996, Berman 1997). This is implicit in democracy's multidimensional character. Good things do not always go together.

Second, some dimensions are undoubtedly more important in guaranteeing a polity's overall level of democracy than others. However, since resolving this issue depends on which overall definition of democracy is adopted and on various causal assumptions that are difficult to prove, the committee is not making judgments on this issue.

Third, it is important to note that dimensions of democracy are not always dimensions of *good governance*. Thus, inclusion of an attribute on this list does not imply that the quality of governance in countries with this attribute will be higher than those without it. For example, some credibly democratic countries (Japan after World War II, the United States in the nineteenth century) have seen enormous corruption scandals. Of course, evaluating whether an attribute of democracy improves the quality of governance hinges on how one chooses to define the latter, about which much has been written but little agreement can be found (Hewitt de Alcantara 1998, Pagden 1998, Knack and Manning 2000). The committee leaves aside the question of how good governance might be defined, noting only that some writers consider democracy an aspect of good governance, some consider good governance an aspect of democracy, and still others prefer to approach these terms as separate and largely independent (nonnested) concepts.

Finally, the committee does not rule out the possibility of alterations to this list of 13. The list might be longer (including additional components) or shorter (involving a consolidation of categories). There is nothing sacrosanct about this particular list of dimensions. Indeed, the committee does not assume that a truly comprehensive set of dimensions is possible, given the extensive and overlapping set of meanings that have been attached to this multivalent term. However, the committee believes strongly that these 13 dimensions are a plausible place to begin.

In any case, whether the index has 13 components or some other (smaller or larger) number is less significant for present purposes than the approach itself. Note that if one begins with a disaggregated set of indicators, it is easy to aggregate upward to create more consolidated concepts. One may also aggregate all the way up to Big D democracy, à la Polity and Freedom House. However, the committee does not propose aggregation rules for this purpose, leaving it as a matter for future scholars and policymakers to decide.

Potential Benefits of Disaggregation

No aggregate democracy index offers a satisfactory scale for purposes of country assessment or for answering general questions pertaining to democracy. Thus, the committee strongly supports USAID's inclination to focus its efforts on a more disaggregated set of indicators as a way of capturing the diverse components of this key concept while overcoming difficulties inherent in measures that attempt to summarize, in a single statistic, a country's level of democracy (à la Freedom House or Polity).

To be sure, before undertaking a venture of this scope and scale, USAID will want to consider carefully the added value that might be delivered by a new set of democracy indicators. In the committee's view, conceptual disaggregation offers multiple advantages. Even so, this approach will not solve every problem, and the committee does not wish to overstate the potential rewards our proposal could bring.

The first advantage to disaggregation is the prospect of identifying concepts on whose definitions and measurements most people can agree. While the world may never agree on whether the overall level of democracy in India can be summarized as a "4" or a "5" (on some imagined scale), it may yet agree on more specific scores along 13 (or so) dimensions for the world's largest democracy.

The importance of creating consensus on these matters can hardly be overemphasized. The purpose of a democracy index is not simply to guide policymakers and policymaking bodies such as USAID, the World Bank, and the International Monetary Fund. Nor could it be so constrained, even if it were desirable. As soon as an index becomes established and begins to influence international policymakers, it also becomes fodder for dispute in other countries around the world. A useful index is one that gains the widest legitimacy. A poor index is one that is perceived as a tool of Western influence or a masque for the forces of globalization (as Freedom House is sometimes regarded). Indeed, because current democracy scales are produced by proprietary scalings and aggregations by specific organizations rather than by objective measurements, those organizations are often subjected to "lobbying" by countries that wish to shift their scores. The hope is that by disaggregating the components of democracy down to levels that are more operational and less debatable, it might be possible to garner a broader consensus around this vexed subject. Countries would know, more precisely, why they received the scores they did. They would also know, more precisely, what areas remained for improvement. Plausibly, such an index might play an important role in the internal politics of countries around the world, akin to the role of Transparency International's Corruption Perceptions Index (Transparency International 2007).

A second advantage is the degree of precision and differentiation that

a disaggregated index offers relative to the old-fashioned “Big D” concept of democracy. Using the committee’s proposed index, a single country’s progress and/or regress could be charted through time, allowing for subtle comparisons that escape the purview of highly aggregated measures such as Freedom House and Polity. One would be able to specify *which facets* of a polity have improved and which have remained stagnant or declined. This means that the longstanding question of regime transitions would be amenable to empirical tests. When a country transitions from autocracy to democracy (or vice versa), which elements come first? Are there common patterns, a finite set of sequences, prerequisites? Or is every transition in some sense, unique?

Similarly, a disaggregated index would allow policymakers to clarify how, specifically, one country’s democratic features differ from others in the region or across regions. While Big D democracy floats hazily over the surface of politics, the dimensions of a disaggregated index are comparatively specific and precise. Contrasts and comparisons may become correspondingly more acute.

Applying the Proposed Index to Democracy Assistance Programming

It is important to remember that, although the committee’s general goal is to provide a path to democracy measures that will be useful to policymakers and citizens alike, the specific charge is to assist USAID. This means the index must be useful for particular policy purposes. Consider the problem of *assessment*. How can policymakers in Washington and in the field missions determine which aspects of a polity are most deficient and therefore in need of assistance? While Freedom House and Polity offer only one or several dimensions of analysis (and these are highly correlated and difficult to distinguish conceptually), the committee’s proposed index would begin with 13 such parameters. It seems clear that for assessing the potential impact of programs focused on different elements of a polity (e.g., rule of law, civil society, governance, and elections—the four subunits of the DG office at USAID), it is helpful to have indicators that offer a differentiated view of the subject.

These same features of the proposed index are equally advantageous for *causal* analysis, which depends on the identification of precise mechanisms, intermediate factors that are often ignored by macro-level cross-national studies. Which aspects of democracy foster (or impede) economic growth? What aspect of democracy is most affected by specific democracy promotion efforts? Whether democracy is looked on as an independent (causal) variable or as a dependent (outcome) variable, we need to know *which* aspect of this complex construct is at play.

Policymakers also wish to know what effect their policy interventions

might have on a given country's quality of democracy (or on a whole set of countries, considered as a sample). There is little hope of answering this question in a definitive fashion if democracy is understood only at a highly aggregated level. The interventions by democracy donors are generally too small relative to the outcome to draw plausible causal inferences between USAID policies, on the one hand, and country A's level of democracy (as measured by Freedom House or Polity) on the other. However, it is plausible—though admittedly still quite difficult—to estimate the causal effects of a project focused on a particular element of democracy if that element can be measured separately. Thus, USAID's election-centered projects might be judged against several specific indicators that measure the characteristics of elections. This is plausible and perhaps quite informative (though, to be sure, many factors other than USAID have an effect on the quality of elections in a country). The bottom line is this: If policymakers cannot avoid reference to country-level outcome indicators, they will be much better served if these indicators are available at a disaggregated meso level.

All of these features should enhance the utility of a disaggregated index for policymakers. Indeed, the need for a differentiated picture of democracy around the world is at least as important for policymakers as it might be for academics. Both are engaged in a common enterprise, an enterprise that has thus far been impeded by the lack of a sufficiently discriminating measurement instrument.

Consider briefly the problem that would arise for macroeconomists, finance ministers, and members of the World Bank and International Monetary Fund if they possessed only one highly aggregated indicator of economic performance. As good as GDP is (and there are, of course, considerable difficulties), it would not go very far without the existence of additional variables that measure the *components* of this macro-level concept. There is a similar situation in the field of political analysis. We have a crude sense of whether countries are democratic, undemocratic, or in between (e.g., "partly free" or partially democratic), but we have no systematic knowledge of how a country should be scored on the various components of democracy.

Since a disaggregated index can be aggregated in a variety of ways, developing a disaggregated index is advantageous even if a single aggregated measure is sometimes desired for policy purposes. Indeed, it is expected that scholars and policymakers will compose summary scores from the underlying data provided by this index. However, the benefit of beginning with the same underlying data (along each of the identified dimensions) is that the process of aggregation is rendered transparent. Any composite index based on these data would be forced to reveal how the summary score for a particular country in a particular year was deter-

mined. Any critic of the proposed score, or of the summary index at large, would be able to contest the aggregation rules used by the author. The methodology is “open source” and thus subject to revision and critique. Further, any causal or descriptive arguments reached on the basis of a summary indicator could be replicated with different aggregation rules. If the results were not robust, it might be concluded that such conclusions were contingent on a particular way of putting together the components of democracy. In short, both policy and scholarly discourse might be much improved by a disaggregated index, *even if* the ultimate objective involves the composition of a highly aggregated index of Big D democracy.

Funding and Management

Readers of this document might wonder why, if the potential benefits of a disaggregated democracy index are so great, one has not yet been developed. There are two simple answers to this question. First, producing such an index would be a time-consuming and expensive proposition, requiring the participation of many researchers. It would not be easy. Second, although the downstream benefits are great, no single scholar or group of scholars has the resources or the incentives to produce such an index.¹⁵ (Academic disciplines do not generally reward members who labor for years to develop new data resources.) Consequently, academics have continued to use—and complain about—Polity, Freedom House, ACLP, and other highly aggregated indices. Policymakers will have to step into this leadership vacuum if they expect the problem of faulty indicators to be solved.

Precedents for such support can be found in other social science fields. USAID served as a principal funder for demographic and health surveys that vastly enhanced knowledge of public health throughout the developing world.¹⁶ The State Department and the Central Intelligence Agency served as principal funders of the Correlates of War data collection project.¹⁷ On a much smaller scale, the State Department provides ongoing support for the Polity project.

To be sure, the entire range of indicators proposed here is probably larger than any single funder is willing or able to undertake. It is highly advisable that several funders share responsibility for the project so that

¹⁵Note that while scholars who are discontented with the leading indicators of democracy periodically recode countries of special concern to them (e.g., McHenry 2000, Berg-Schlosser 2004a, b; Acuna-Alfaro 2005; Bowman et al 2005), this recoding is generally limited to a small set of countries and/or a small period of time.

¹⁶Surveys and findings are described on the USAID Web site: <http://www.measuredhs.com/>.

¹⁷Information about the project may be found at <http://www.correlatesofwar.org/>.

its financial base is secure into the future and so that the project is not wholly indebted to a single funder, a situation that might raise questions about project independence. Preferably, some of these funders would be non-American (e.g., Canadian, European, Japanese, European Union, or international organizations like the World Bank or the United Nations Development Program). Private foundations (e.g., Open Society Institute, Google Foundation) might also be tapped. The committee conceptualizes this project as a union of many forces. This makes project management inevitably more complicated. However, the sorts of difficulties encountered here, insofar as they constitute a deliberative process about the substantive issues at stake, may enhance the value of the resulting product. Certainly, it will enhance its legitimacy.

Another possibility is that different funders might undertake to develop (or take responsibility for) different dimensions of the index, thus apportioning responsibility. It is preferable, in any case, that some level of supervision be maintained at the top so that the efforts are well coordinated. Coordination involves not only logistical issues (sharing experts in the field, software, and so forth) but also, more importantly, the development of indicators that are mutually exclusive (nonoverlapping) so that the original purpose of the project—disaggregation—is maintained. Note that several of the above-listed components might be employed across several dimensions, requiring coordination on the definition and collection of that variable.

As a management structure, the committee proposes an advisory group to be headed by academics—with some remuneration, depending on the time requirements, and suitable administrative support—in partnership with the policy community.¹⁸ This partnership is crucial, for any widely used democracy assessment tool should have *both* a high degree of academic credibility *and* legitimacy among policymakers. Major shortcomings of previous efforts to develop indices of democracy and governance resulted from insufficient input from methodologists and subject specialists or lack of broad collaboration across different stakeholders.

For this wide-ranging proposal, experts on each of the identified dimensions will be needed. Their ongoing engagement is essential to the success of the enterprise. Moreover, it is important to solicit help widely within the social sciences disciplines so that decisions are not monopolized by a few (with perhaps quirky judgments). As a convening body,

¹⁸The Utstein Partnership, a group formed in 1999 by the ministers of international development from the Netherlands, Germany, Norway, and the United Kingdom to formalize their cooperation is an example of this possible approach applied to a different problem. The U4 Anti-Corruption Resource Centre assists donor practitioners to more effectively address corruption challenges by providing a variety of online resources. See <http://www.u4.no/about/u4partnership.cfm>.

there are several possibilities, including the professional associations of political science, economics, and sociology (the American Political Science Association, American Economic Association, and American Sociological Association) or a consortium of universities.

CONCLUSIONS

This chapter has reviewed the most widely used indicators that measure “democracy” and arrived at these key findings:

- The concept of democracy cannot at present be defined in an authoritative (nonarbitrary) and operational fashion. It is an inherently multidimensional concept, and there is little consensus over its attributes. Definitions range from minimal—a country must choose its leaders through contested elections—to maximal—a country must have universal suffrage, accountable and limited government, sound and fair justice and extensive protection of human rights and political liberties, and economic and social policies that meet popular needs. Moreover, the definition of democracy is itself a moving target; definitions that would have seemed reasonable at one time (such as describing the United States as a democracy in 1900 despite no suffrage for women and few minorities holding office) are no longer considered reasonable today. To obtain a more reliable and credible method of tracking democratic change to guide USAID DG programming, USAID should foster an effort to develop disaggregated sectoral-level measures of democratic governance. This would likely have to involve numerous parties to attain wide acceptance.

- Existing empirical indicators of democracy are flawed. The flaws extend to problems of definition and aggregation, imprecision, measurement errors, poor data coverage, and a lack of convergent validity. These existing measures are useful to identify whether countries are fully democratic, fully autocratic, or somewhere in between. They are not reliable, however, as a guide for tracking modest improvements or declines in democracy within a country over the period of time in which most DG projects operate.

- While the United States, other donor governments, and international agencies that are making decisions about policy in the areas of health or economic assistance are able to draw on extensive databases that are compiled and updated at substantial cost by government or multilateral agencies mandated to collect such data (e.g., World Bank, World Health Organization, Organization for Economic Cooperation and Development), no comparable source of data on democracy currently exists. Data on democracy are instead currently compiled by various individual academics on irregular and shoestring budgets, or by nongovernmental

organizations or commercial publishers, using different definitions and indicators of democracy.

These findings lead the committee to make a recommendation that we believe would significantly improve USAID's (and others') ability to track countries' progress and make the type of strategic assessments that will be most helpful for DG programming.

- **USAID and other policymakers should explore making a substantial investment in the systematic collection of democracy indicators at a disaggregated, sectoral level—focused on the components of democracy rather than (or in addition to) the overall concept.** If they wish to have access to data on democracy and democratization comparable to that relied on by policymakers and foreign assistance agencies in the areas of public health or trade and finance, a substantial government or multilateral effort to improve, develop, and maintain international data on levels and detailed aspects of democracy would be needed. This should not only involve multiple agencies and actors in efforts to initially develop a widely accepted set of sectoral data on democracy and democratic development but should seek to institutionalize the collection and updating of democracy data for a broad clientele, along the lines of the economic, demographic, and trade data collected by the World Bank, United Nations, and International Monetary Fund.

While creating better measures at the sectoral level to track democratic change is a long-term process, there is no need to wait on such measures to determine the impact of USAID's DG projects. USAID has already compiled an extensive collection of policy-relevant indicators to track specific changes in government institutions or citizen behavior, such as levels of corruption, levels of participation in local and national decision making, quality of elections, professional level of judges or legislators, or the accountability of the chief executive. **Since these are, in fact, the policy-relevant outcomes that are most plausibly affected by DG projects, the committee recommends that measurement of these factors rather than sectoral-level changes be used to determine whether the projects are having a significant impact in the various elements that compose democratic governance.**

REFERENCES

- Acuna-Alfaro, J. 2005. *Measuring Democracy in Latin America (1972-2002)*. Working Paper No. 5, Committee on Concepts and Methods (C&M) of the International Political Science Association. Mexico City: Centro de Investigacion y Docencia Economicas.

- Adcock, R., and Collier, D. 2001. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review* 95(3):529-546.
- Altman, D., and Pérez-Liñán, A. 2002. Assessing the Quality of Democracy: Freedom, Competitiveness and Participation in Eighteen Latin American Countries. *Democratization* 9(2):85-100.
- Alvarez, M., Cheibub, J.A., Limongi, F., and Przeworski, A. 1996. Classifying Political Regimes. *Studies in Comparative International Development* 31(2):3-36.
- Arat, Z.F. 1991. *Democracy and Human Rights in Developing Countries*. Boulder: Lynne Rienner.
- Beck, T., Clarke, G., Groff, A., Keefer P., and Walsh, P. 2000. New Tools and New Tests in Comparative Political Economy: The Database of Political Institutions. Policy Research Working Paper 2283. Washington, DC: World Bank, Development Research Group. For further info, see <http://www.worldbank.org/research/bios/pkeefee.htm> Research Group Web site <http://econ.worldbank.org/>.
- Beetham, D., ed. 1994. *Defining and Measuring Democracy*. London: Sage.
- Berg-Schlosser, D. 2004a. Indicators of Democracy and Good Governance as Measures of the Quality of Democracy in Africa: A Critical Appraisal. *Acta Politica* 39(3):248-278.
- Berg-Schlosser, D. 2004b. The Quality of Democracies in Europe as Measured by Current Indicators of Democratization and Good Governance. *Journal of Communist Studies and Transition Politics* 20(1):28-55.
- Berman, S. 1997. Civil Society and the Collapse of the Weimar Republic. *World Politics* 49(3): 401-429.
- Bertelsmann Foundation. 2003 *Bertelsmann Transformation Index: Towards Democracy and a Market Economy*. Gütersloh, Germany: Bertelsmann Foundation.
- Boix, C., and Rosato, S. 2001. *A Complete Data Set of Political Regimes, 1800-1999*. Chicago: University of Chicago, Department of Political Science.
- Bollen, K.A. 1980. Issues in the Comparative Measurement of Political Democracy. *American Sociological Review* 45:370-390.
- Bollen, K.A. 1993. Liberal Democracy: Validity and Method Factors in Cross-National Measures. *American Journal of Political Science* 37(4):1207-1230.
- Bollen, K.A., and Paxton, P. 2000. Subjective Measures of Liberal Democracy. *Comparative Political Studies* 33(1):58-86.
- Bowman, K., Lehoucq, F., and Mahoney, J. 2005. Measuring Political Democracy: Case Expertise, Data Adequacy, and Central America. *Comparative Political Studies* 38(8):939-970.
- Casper, G., and Tufis, C. 2003. Correlation Versus Interchangeability: The Limited Robustness of Empirical Findings on Democracy Using Highly Correlated Data Sets. *Political Analysis* 11:196-203.
- Coppedge, M. 2007. Presentation to *Democracy Indicators for Democracy Assistance*. Boston University, January 26.
- Coppedge, M. Forthcoming. *Approaching Democracy*. Cambridge: Cambridge University Press.
- Coppedge, M., and Reinicke, W.H. 1990. Measuring Polyarchy. *Studies in Comparative International Development* 25:51-72.
- Dunn, J. 2006. *Democracy: A History*. New York: Atlantic Monthly Press.
- Elkins, Z. 2000. Gradations of Democracy? Empirical Tests of Alternative Conceptualizations. *American Journal of Political Science* 44(2):287-294.
- Finkel, S.E., Pérez-Liñán, A., and Seligson, M.A. 2007. The Effects of U.S. Foreign Assistance on Democracy Building, 1990-2003. *World Politics* 59(3):404-439.
- Finkel, S.E., Pérez-Liñán, A., Seligson, M.A., and Tate, C.N. 2008. Deepening Our Understanding of the Effects of U.S. Foreign Assistance on Democracy Building: Final Report. Available at: <http://www.LapopSurveys.org>.

- Foweraker, J., and Krznaric, R. 2000. Measuring Liberal Democratic Performance: An Empirical and Conceptual Critique. *Political Studies* 48(4):759-787.
- Freedom House. 2006. *Freedom of the Press 2006: A Global Survey of Media Independence*. New York: Freedom House.
- Freedom House. 2007. *Methodology, Freedom in the World 2007*. Freedom House: New York. Available at: http://www.freedomhouse.org/template.cfm?page=351&ana_page=333&year=2007. Accessed on September 5, 2007.
- Gasiorowski, M.J. 1996. An Overview of the Political Regime Change Dataset. *Comparative Political Studies* 29(4):469-483.
- Gleditsch, K.S., and Ward, M.D. 1997. Double Take: A Re-examination of Democracy and Autocracy in Modern Polities. *Journal of Conflict Resolution* 41:361-383.
- Hadenius, A. 1992. *Democracy and Development*. Cambridge: Cambridge University Press.
- Hadenius, A., and Teorell, J. 2005. Assessing Alternative Indices of Democracy. Committee on Concepts and Methods Working Paper Series. Mexico City: Centro de Investigacion y Docencia Economicas (CIDE).
- Hewitt de Alcantara, C. 1998. Uses and Abuses of the Concept of Governance. *International Social Science Journal* 11(155):105-113.
- Kaufmann, D., Kraay, A., and Mastruzzi, M. 2006. *Governance Matters V: Governance Indicators for 1996-2005*. Washington, DC: World Bank.
- Kecic, L. 2007. The Economist Intelligence Unit's Index of Democracy. Available at: http://www.economist.com/media/pdf/DEMOCRACY_INDEX_2007_v3.pdf. Accessed on February 23, 2008.
- Knack, S., and Manning, N. 2000. *Why Is It So Difficult to Agree on Governance Indicators?* Washington, DC: World Bank.
- Kurtz, M.J., and Schrank, A. 2007. Growth and Governance: Models, Measures, and Mechanisms. *Journal of Politics* 69:2.
- Landman, T. 2003. Map-Making and Analysis of the Main International Initiatives on Developing Indicators on Democracy and Good Governance. Unpublished manuscript, University of Essex.
- Levi, M. 1996. Social and Unsocial Capital: A Review Essay of Robert Putnam's Making Democracy Work. *Politics & Society* 24:145-155.
- McHenry, D.E. 2000. Quantitative Measures of Democracy in Africa: An Assessment. *Democratization* 7(2):168-185.
- Moon, B.E., Birdsall, J.H., Ciesluk, S., Garlett, L.M., Hermias, J.H., Mendenhall, E., Schmid, P.D., and Wong, W.H. 2006. Voting Counts: Participation in the Measurement of Democracy. *Studies in Comparative International Development* 41(2):3-32.
- Munck, G.L. 2006. Standards for Evaluating Electoral Processes by OAS Election Observation Missions. Paper prepared for Organization of American States.
- Munck, G.L., and Verkuilen, J. 2002. Conceptualizing and Measuring Democracy: Alternative Indices. *Comparative Political Studies* 35(1):5-34.
- Pagden, A. 1998. The Genesis of Governance and Enlightenment Conceptions of the Cosmopolitan World Order. *International Social Science Journal* 50(1):7-15.
- Reich, G. 2002. Categorizing Political Regimes: New Data for Old Problems. *Democratization* 9(4):1-24.
- Schaffer, F.C. 1998. *Democracy in Translation: Understanding Politics in an Unfamiliar Culture*. Ithaca, NY: Cornell University Press.
- Schumpeter, J.A. 1942. *Socialism and Democracy*. New York: Harper.
- Tetlock, P. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.
- Thomas, M.A. 2007. What Do the Worldwide Governance Indicators Measure? Unpublished manuscript, School of Advanced International Studies, Johns Hopkins University.

- Transparency International. 2007. Corruption Perceptions Index. Available at: http://www.transparency.org/policy_research/surveys_indices/cpi. Accessed on September 5, 2007.
- Treier, S., and Jackman, S. 2003. Democracy as a Latent Variable. Paper presented at the Political Methodology meetings, University of Minnesota, Minneapolis-St. Paul.
- USAID (U.S. Agency for International Development). 1998. *Handbook of Democracy and Governance Program Indicators*. Washington, DC: Center for Democracy and Governance. Available at: http://www.usaid.gov/our_work/democracy_and_governance/publications/pdfs/pnacc390.pdf. Accessed on August 1, 2007.
- Vanhanen, T. 2000. A New Dataset for Measuring Democracy, 1810-1998. *Journal of Peace Research* 37:251-265.
- Vermillion, J. 2006. Problems in the Measurement of Democracy. *Democracy at Large* 3(1): 26-30.

Learning from the Past: Using Case Studies of Democratic Transitions to Inform Democracy Assistance

INTRODUCTION

The U.S. Agency for International Development (USAID) asked the National Research Council (NRC) to recommend methodologies to carry out retrospective analyses of democracy assistance programs. The recommendations were to include “a plan for cross-national case-study research to determine program effectiveness and inform strategic planning.”

There is a substantial and growing literature of case studies of democracy assistance programs, many of them commissioned by USAID or other agencies engaged in democracy assistance. The goal of such case studies is to learn what has worked and what has not among the varied democracy and governance (DG) programs in a variety of places.

The vast majority of such studies focus on a particular program in a particular country, such as human rights in Cambodia (Asia Watch 2002), party organization in Uganda (Barya et al 2004), voter education in Ethiopia (McMahon et al 2004), or justice reform in Sierra Leone (Dougherty 2004).

In addition, there have been more ambitious works that looked at multiple countries to try to draw broader lessons about program impacts. For example, Abbink and Hesselning (2000) bring together several studies of election observation and democratization in Africa; Lippman and Emmert (1997) study legislative assistance in five countries; Blair and Hansen (1994) assess the impact of rule of law programs in six countries; Kumar (1998) examines the impact of elections in several postconflict conditions; O’Neill (2003) presents lessons from human rights promotion in

varied regions; Carter et al (2003) study the overall impact of USAID DG programs in six countries; and de Zeeuw and Kumar (2006) look at media, human rights, and election programs in nine postconflict states.

While these studies have generated valuable insights into how programs were carried out, how they were received, and how participants and donors perceived their effects, they are not ideal either for “determining program effectiveness” or to “inform strategic planning.” This is because such studies focused almost entirely on specific DG projects, rather than on the broader context of democratization in the countries being studied. They did not systematically compare cases of varying levels of DG assistance or compare the effects of DG projects with comparison groups that did not receive assistance.

CASE STUDY DESIGNS AND METHODS

The basic tool of case study analysis is process tracing (George and Bennett 2005). In this method, researchers track the unfolding of strings of events, testing hypotheses regarding the causal relationships among them by considering multiple hypotheses that could account for the strings of events and searching for confirming and disconfirming evidence. The process is not unlike a detective’s efforts to solve a murder mystery by reconstructing a timeline of events, examining all possible suspects and their alibis, assessing plausible motives and opportunities for the observed actions and events, and building a case in favor of one causal chain as having determined the ultimate outcome rather than others.

Like solving any mystery, process tracing can be painstaking and time-consuming work, and the results often depend on an analyst’s skill in recognizing how specific social conditions, motivations, events, and opportunities link to form a coherent explanatory chain. Also like any criminal case, the persuasiveness of pointing out any one factor or event as causal depends on the analyst’s imagination and skill in identifying and considering *alternative causal pathways* and gathering evidence as to how likely or unlikely they were.¹

Case studies to demonstrate the effectiveness of aid programs thus face the same challenge as formal statistical evaluations—they must try to determine what would have happened in the absence of the aid program, whether by including studies of both groups receiving aid and those not receiving aid in their case studies (a comparative case study design) or by trying to trace and account for historical trends and confounding fac-

¹Hence the famous quote from Sherlock Holmes in *Adventure of the Beryl Coronet*: “When you have excluded the impossible, whatever remains, however improbable, must be the truth” (Doyle 1998).

tors to estimate the likely causal chains that would have unfolded in the absence of the aid program (a long-term historical case study design).

Yet in most case studies of democracy assistance, researchers have not used such designs. They have instead assumed that the information they needed could be found by studying the unfolding of the aid program itself. For a process-type evaluation, where the main questions asked by researchers are “Did the project achieve the goals expected by the donors?” and “Why or why not?” this is reasonable and most case studies of aid assistance have taken this form.

However, if USAID now wishes to use case studies to study the impact of DG programs on policy goals, they are not the most appropriate tool. This is because retrospective case studies can rarely obtain or reconstruct the comparable baseline and outcome information for appropriate comparison groups that is necessary for sound inference of program effects. The committee’s field studies tried to determine if missions had retained such baseline data if collected before DG projects or if they had collected any comparable baseline data for nonassisted groups. The teams had limited success with finding the former and no success in finding the latter. Thus the committee believes that for most DG programs information on project effects would most credibly be obtained by well-designed impact evaluations, rather than retrospective case studies.

However, case studies can provide information to help inform strategic planning. Comparative and historical case studies that examine varied trajectories of democratic change, and trace the relationship of DG activities to other factors and events that influence long-term democracy outcomes, can help generate hypotheses about opportunities and obstacles for DG assistance to support democratic progress.

In addition, sometimes the greatest insights regarding where and when to intervene with certain programs arise from detailed studies of program failures. One can often learn more from tracing the causes of program failure than from studies of successes, especially if such success rests on chance factors that supported a program but are not observed or reported in the study. Yet case studies of DG assistance rarely seek out failures for sustained examination—there are few rewards in the current incentive structure of donors for seeking out failures and investing in their study.²

This chapter develops guidelines for case studies that better explore the roles that democracy assistance programs may play in varied contexts of social change.

²One exception is the scholarly work of Carothers (1999, 2004, 2006), who has investigated instances of disappointing results in democracy assistance programs.

INSIGHTS FROM CURRENT RESEARCH: RESULTS OF A CONFERENCE OF CASE STUDY SPECIALISTS ON DEMOCRACY

Under the “transformational diplomacy” plan of the Bush administration and the closer supervision of USAID by the State Department, it was anticipated that USAID’s DG efforts would often be undertaken as part of broader strategies to help achieve desired outcomes in particular states (Rice 2006). Faced with such demands, USAID would like to be able to respond to policymakers with information such as the following: “Based on what we know about transitions to democracy in countries with conditions like that, the chances of achieving a successful transition to democracy in X years is fairly low (or high),” or “Based on what we know about the time and volume of assistance it usually takes to build and stabilize democracy in postconflict societies with these characteristics, we can give you some broad parameters regarding the expected time and financial support required to have a realistic chance of attaining that goal in country Y.”

For these objectives a clustered set of case studies, tracing the processes through which advances toward democracy were made from various sets of initial conditions, is an appropriate mode of investigation. A sufficient number of case studies would help build a knowledge base to answer questions such as the following: “For most countries we have observed with initial conditions X, Y, and Z, what have been the observed trajectories of political change, and which factors A, B, C (and others) were most prominent in shaping or constraining those trajectories?”

Case studies are particularly valuable in this kind of mapping exercise, where instead of trying to identify the average impact of one or more causal factors across a wide range of conditions, the goal of the investigation is to identify diverse patterns or combinations of relationships that are associated with varying pathways of change over time (Goldstone 1998, 2003).

Rather than starting out to design such a study, the committee first noted that a great deal of case study research is already being done by academics who focus on democracy and democratization. The committee decided that its first step should be to investigate that body of scholarship and see how much value it already provided for meeting USAID’s goals. The committee therefore convened a conference of leading academic experts on case study analyses of democracies and democratic transitions to help it assess the “state of the art” on how such knowledge could guide strategies for democracy assistance (see Appendix D for the details of this conference).

This section presents the main findings that emerged during that conference, followed by the committee’s own conclusions and recommendations for future studies. The committee does not present the fol-

lowing findings as definitive, nor are they endorsed as the results of the committee's own research. Rather, what follows is a synopsis of the main points expressed by scholars at the conference, with particular attention to findings relevant to either DG assistance planning or research designs for case studies of DG assistance programs.

I. Democracy research conducted by the academic community generally needs considerable translation to be useful for guiding democracy assistance.

One problem that was immediately evident from discussions between the scholars and practitioners who attended the conference is that much of the academic research on democracy and democratic transitions is not developed or presented in ways that offer much practical guidance to policy professionals. This is much more than a simple matter of pure versus applied research. Rather, policymakers dealing with democracy assistance simply have to act in much more constrained circumstances than the typical academic study implies.

For example, Terry Karl of Stanford University noted that one major conclusion of her research was that agreements, which she terms "pacts," should be developed among elites *before* elections, rather than holding elections first and hoping to bring agreements among elites afterward. As an academic finding, this seems impeccable—an increasingly large body of empirical and theoretical work argues that elections can be stabilizing if they affirm agreements that bridge social cleavages and unite diverse elites in a commitment to abide by democratic rules, but tend to be *destabilizing* if the elections harden or polarize prior social cleavages and pit rival elites against each other in a zero-sum struggle for control of society (Berman 1997, 2001; Goldstone and Ulfelder 2004; Zakaria 2004; cf. the election in Kenya in December 2007).

However, the reality facing policymakers is that they are often called on to organize and hold elections that are demanded by the society in question, or by the international community, in which influential and critical actors are not prepared to wait until after a pact has been agreed on (Carothers 2007). Unless the weight of experience and academic research reduce the current pressures felt by policymakers to hold elections as soon as possible in emerging democracies or postconflict states, some group needs to take up the challenge of translating the findings of academic research into guidelines for actions that can be more flexible and adapted to adverse or rapidly changing conditions. Thus, one lesson to draw from Professor Karl's research may be that when elections need to be held rapidly in the absence of prior pacts, the electoral process should be designed as much as possible to lead rival factions to seek pacts in the process of seeking electoral success. That is, rules on the composition of electoral

commissions, or restrictions on parties to require party lists to have cross-group representation, or voting schemes that require regionally dispersed support to attain electoral success should be developed to use the election process itself to bring elites together and to “tame” factionalism.

While the specific adjustments must be tailored to each case (from using an extant body with strong legitimacy that has traditionally bridged factions, like the Afghan *loya jirga*, as part of the process, to the requirements in Nigeria and Kenya that candidates demonstrate cross-regional support to qualify for the ballot), the translation process needs to show how a clear but academic principle—“pacts before elections”—can be adapted to the rough-and-tumble and uncontrolled circumstances of actual transition policymaking and response.

One finding from the conference was thus that, although a large number of meetings between academics and policy professionals do occur (e.g., under the aegis of the National Endowment for Democracy), a more structured forum in which policymakers and academics can spend time focusing on discussing one particular type of policy intervention, or one group of countries, is needed if academics and policy professionals are to become able to understand each other fully and gain from each other’s knowledge and experience. It often appeared in the committee’s meeting that academics were interested in offering broad general insights or developing abstract categories to sort out developments in a large number of states, while policy professionals worried more about what would help them deal with the rapidly changing conditions and diverse pressures they face on the job.

To answer the question “How do you best assist the development of democracy under these conditions?” academic researchers and policy professionals first need to work out some agreement on what they consider to be the relevant conditions. Where academics usually will define them by abstract or historical categories, policy professionals will more often refer to the conditions under which they are expected to work. A host of such issues of varying vocabulary and references need to be worked out by direct communications before the fruits of academic research are likely to answer questions posed by USAID professionals and vice versa.

II. Democracy assistance donors and policymakers need to be aware that donors do not control the context.

In approaching the question “How much time and resources will it typically take to help secure a democratic outcome in a country like X?” it became clear that this query is not phrased correctly. This is because, as the academic scholars repeatedly noted and the practitioners readily acknowledged, democracy assistance providers do not control the context

in which they work. Thus it is not always possible to form stable estimates of the likelihood or costs of attaining specific outcomes.

First, this principle means that expectations for success in democracy assistance must be tempered. A host of issues impinge on a country's progress toward democracy—for example, standards of living, government structures, international influences, regional conditions—that are usually completely beyond the ability of democracy assistance donors to affect.³ Thus democracy assistance always needs to be opportunistic as well as strategic, identifying promising steps that can be taken in both the short term and the long term and then being ready to assist when conditions rapidly change and new openings for democracy arise.

Second, because context is more generally controlled and opportunities more readily grasped by members of the society than by outsiders, democracy assistance is only effective when supporting the activity of committed individuals and groups *within* the society and cannot be successfully manipulated wholly from the outside. This point is often made by those with experience in democracy assistance, such as de Zeeuw and Kumar (2006:282): “Although external actors can perhaps do more to avoid legitimating political window-dressing and thwart the incentives for corrupt activities, in the end it is up to domestic political leaders to stop these practices.”⁴

Third, the inability to control context means that the success of democracy assistance efforts can rarely be judged in the short term with regard to overall progress toward democracy. Rather, such success has to be judged in terms of whether any steps that may contribute to future democracy are leaving a demonstrable footprint on institutions or behavior; whether reactions to opportunities were prompt, creative, and effective in using such opportunities to assist democratic reformers and efforts to secure democracy; and whether steps that reverse democratic progress are being discouraged. Modest success in the face of the most discouraging and hostile contexts is a considerable achievement, while being able to take advantage of the most favorable contexts is probably the most cost-effective approach to improving democratic prospects.

Given that context varies greatly and that many elements important to

³Although there is much debate on the conditions that facilitate democratic transitions and consolidation, empirical work by Barro (1999), Boix and Stokes (2003), and Epstein et al (2006) all concur that economic performance is a major factor in democratic transitions, while studies by Haggard and Kauffman (1995) and Przeworski et al (2000) underline the importance of economic performance for democratic consolidation. Goldstone and Ulfelder (2004) also point to the importance of such factors as the presence of ethnic or religious discrimination and conflicts in neighboring countries as key factors that can undermine democracies.

⁴This point is also emphasized by Dobbins (2003) and McFaul (2006).

democratic development in a society are beyond the control of democracy assistance donors, it is probably wrong to ask “how much” time, effort, or expense will be required to “move” a country into the democratic column. More realistically, it could be asked under what conditions might what kind of investments pay off and in what time frame?

This also has implications for framing any case studies of democracy assistance. Given the vital importance of widely varying contexts, case studies would need to cover a substantial range of contexts that favor or disfavor democratization, not merely a diverse set of nations.

III. Democratic transitions are highly nonlinear processes.

A linear process is one that occurs in a fairly smooth and continuous fashion and in which outputs change in proportion to various inputs. Unfortunately, democratic transitions do *not* have this character. Instead, such transitions are often sudden and discontinuous events, in which little or no change is observed at the national level for a long time, and then rapid shifts in power or political conditions occur. Similarly, even emerging democracies that appear to be stable can suddenly be overturned by an antidemocratic coup (e.g., Thailand) or collapse into violent conflicts (Nepal, Rwanda, Côte d’Ivoire).⁵

This nonlinearity has major implications for planning and assessing democracy assistance policies. It means that the impact of democracy assistance in a given nation cannot simply be measured by looking for a smooth and proportional movement to democracy in response to such assistance. Instead, it may take years for the impact of democracy assistance to be revealed in the course of a sudden transition.

For example, in a recent study of the democratic transition in the Ukraine, McFaul (2006) argues that during many years of President Leonid Kuchma’s regime, democracy assistance aimed at strengthening the media, improving the autonomy of the judiciary, upgrading election commissions, and building civil society and party organizations had little or no impact on the nature of Ukraine’s regime. However, when an opening for democratic action arose during the maneuvering around elections to choose Kuchma’s successor, particularly around suspicions that the elections were fraudulent, the institutions that had been strengthened by external democracy assistance helped challenge the efforts of the Kuchma regime to control the electoral outcome. McFaul’s analysis concludes that the impact of democracy assistance was thus only “revealed” when new opportunities arose for challenging the authoritarian regime.

⁵For a detailed examination of nations’ trajectories toward democracy since World War II, which illustrates how “bouncy” and “jerky” such transitions have been, see Goldstone (2007).

This nonlinearity also reinforces the point made above that democracy assistance itself must be flexible, patient, and opportunistic. Furthermore, when transitions occur, they cannot be taken for granted as having achieved a new and therefore stable equilibrium. Rather, aid may need to be sustained and retargeted to support emerging democracies for a considerable period in order to hold off sudden backsliding or collapse or to respond to new threats to democratic stability.

This nonlinearity also has major implications for the conduct of research on the impact of democracy assistance. Rather than looking for the impact of such assistance simply by focusing on the area receiving aid and searching for near-term impacts, it is necessary to place such assistance in a longer term and large-scale context. While the specific forms of assistance need to be related to changes in the character of specific institutions or behaviors, researchers must then address the full process of democratic change, sustainability, or retreat over a considerable period in the country where assistance is being studied in order to identify lagging and late-emerging effects. Without attention to the impact of contingency and changing context on a longer scale, a full and accurate assessment of democracy assistance is unlikely.

IV. Different policy guidelines are needed for different democratization contexts.

The scholars at the Stanford conference identified at least three distinctive contexts in which donors have been active in providing democracy assistance: (1) currently authoritarian and semiauthoritarian regimes, (2) transition and posttransition regimes, and (3) postconflict regimes. Recognizing that there can be many arguments over how to categorize regimes—and even what categories to use—they suggested that these three offer particular opportunities and constraints for democracy assistance.

A. Authoritarian and semiauthoritarian regimes

Authoritarian regimes are those in which a single individual or group (e.g., a single party or the military) wields unshakable power. There may be greater or lesser subordinate powers, even some with a democratic façade (e.g., elected but pliant legislatures, subordinate parties with no chance of acquiring power), but there is no question where ultimate decision-making power resides and that authority faces no effective checks or accountability. Under such conditions, as long as the authoritarian regime has sufficient resources and elite support, only incremental progress toward building the foundations of democracy is possible. The scholars suggested that useful actions could include promoting transparency in government finance, fighting corruption, and promoting human rights. The goal of these actions is to seek to open a space in which the

absolute authority of the leadership can be subjected to scrutiny or criticism. Engagement in international relations, including trade, educational exchange, diplomatic relations, and information/broadcasting, is useful for providing leverage and openings for these causes, which are almost impossible to advance solely from outside in the absence of any relations with the country. Support for democrats within the society—insofar as can be done without undermining their legitimacy by making them appear as subordinate to external powers—also can help advance the foundations for future democratic reform.

Many scholars insisted on a further distinction between “hard” authoritarian regimes, also called “full autocracies,” in which all opposition is ruthlessly crushed and dissent is not tolerated (as in Saddam’s Iraq or Stalin’s Soviet Union), and “semiauthoritarian regimes” (also called “partial autocracies”). In these latter regimes, power is still monopolized by a single person or group. However, there are also limited openings for opposition to appear. There may be some press or media outlets that are independent of the regime; there may be opposition parties that, while small and ineffective, are not co-opted or repressed by the ruler; there may be professional organizations or even some elements of government—certain judges or commissions—that operate autonomously and have some respect and authority apart from their support of the regime. Examples include the Philippines under Ferdinand Marcos, Nicaragua under the Somozas, and Ukraine under Kuchma. Several studies—both using case studies (McFaul 2005) and large *N* statistical analysis (Epstein et al 2006)—have argued that such partial autocracies are more likely to make the move to democratic politics than are full autocracies.

In the authoritarian context, major advances toward democracy are usually dependent on crisis events that weaken the regime but that democracy assistance donors cannot create or control. Typically, such crises include war, fiscal or monetary collapse, a looming succession, exposure of corruption, a major repressive overstep by the regime, natural disasters, or an electoral surprise (e.g., unexpected results in an election that would normally be fully controlled by the regime). Such events create a window of opportunity in which democracy assistance has the chance to be more powerful. In the wake of such events, democracy assistance that would be infeasible or ineffective under a firm authoritarian regime, such as support for opposition organizations, support for independent media, or support for election monitors/commissions, may help local democratic forces use the opportunity to press for major reforms.

B. Transition and posttransition regimes

Transition and posttransition regimes are those in which a democratic regime has been established but has not yet been consolidated by repeated

peaceful and effective electoral choice of leaders and the secure institutionalization of civil and political freedoms. In this context, a relatively long-term commitment to the support and improvement of democratic behavior and institutions may promote democratic stability.

The scholars cited one major problem of democracy assistance in this context: External assistance often is increased in authoritarian contexts, or at the time of transition, but then swiftly reduced after the initial transition to democracy. They argued that instead a steady flow of assistance through a substantial posttransition period is often needed to help stabilize the new democracy and avoid backsliding or to head off subsequent crises.

The list of actions needed to support democratic stabilization is lengthy, for during this period many aspects of democratic institutions may need nurturing or protection, and the society is relatively open to receiving such support. Areas that might benefit from assistance in this phase include assuring the competitiveness of multiple political parties that are inclusive and able to compromise; consolidating free, fair, and inclusive electoral procedures; developing legislatures that are effective in writing and passing needed legislation; improving the accountability of government at national and local levels; supporting varied media; promoting transparency, human rights, and fighting corruption; building a fair and effective criminal and civil justice system (police and judiciary); establishing a professional military that is subordinate to civilian control; improving social services (health, education, sanitation); and improving economic performance. Careful assessment is needed to determine which donors and agencies are best suited to assist in these varied areas, which areas require the most help, and whether such commitments can be sustained.

C. *Postconflict regimes*

Postconflict regimes are those in which recent conflict has left either an absence of central political authority or a weak central authority unable to control violence and crime or unable to control local warlords or suppress regional rebellions. There may be an authoritarian or democratic regime trying to acquire power over the society or the country may be divided, with various regions held by conflicting groups, warlords, or rebels.

For postconflict regimes several of the scholars at the conference pointed to a smaller number of key tasks that are imperative to complete if further actions to help achieve democracy are to have a chance of success. These were (1) reduce factional conflicts by building elite cooperation and agreements; (2) create security by establishing military and policy protection of civilians by the central regime and undertaking dis-

armament of rebels, warlords, or other competing authorities; (3) design and secure agreement on constitutional and electoral processes that will promote inclusion, participation, and legitimacy for the regime; (4) create effective processes for the integration of combatant and extremist groups into civilian society; and (5) create truth and reconciliation processes that will blunt the drive for personal and arbitrary vengeance. If these steps are not successfully completed, other steps—such as building political parties or holding elections—are unlikely to bear fruit, and conflict is likely to recur. One of the problems of democracy assistance programs in places such as Iraq or other postconflict contexts has been a tendency by donors to jump to the activities listed under B above without first achieving the five items listed here for postconflict regimes. Yet without making substantial progress on most or all of these five items, efforts on the activities listed under B are not likely to be effective in helping to achieve democracy in postconflict settings.

It is crucial to realize that the above comments represent rather sweeping but preliminary generalizations from current academic research on democracy. There are, in fact, a variety of kinds of authoritarian and semi-authoritarian regimes, ranging from hereditary monarchies and military dictatorships to one-party states, and similarly a variety of postconflict conditions depending on the nature, severity, and extent of the conflict. The broad goals cited above for various contexts also still leave as highly problematic whether, and which, specific actions have significant effects in advancing those goals. Thus the only true conclusion at this point is that context matters greatly, both for designing policy and for planning future research on democratization and democracy assistance.

V. Popular protest and mobilization are a double-edged sword.

Democracy assistance donors often face very difficult choices regarding popular protest and mobilization. Should change be pursued by encouraging popular protest or only through formal and institutional means? Should one work mainly through elites, or is it better to pressure or outflank elites through popular movements? If popular movements are currently mobilizing or a protest wave is starting in a currently authoritarian state or transition state, should it be encouraged, viewed as an opportunity to push further change, or blunted as a potential threat to creating dangerous disorder?

The scholars at the Stanford conference suggested that popular protest is often an important factor in encouraging democratic transitions but noted that mobilization needs to be diverted into electoral activity and civil society organizations—rather than militias, populist movements, or competing factions—if democratic consolidation is to occur.

Popular protests have frequently played a crucial role in turning crises of opportunity into democratic transitions. Protest—or fear of pro-

test—often forces weak leaders to abandon office and forces elites to enter into pacting agreements. These are positive elements in the development of democracy from authoritarian regimes.

However, it is imperative that inclusive and effective political parties emerge to channel popular mobilization into peaceful political competition. Otherwise, popular groups may be mobilized into support for ethnic or regional groups, individual populist leaders, or even militias that become major security threats. In such cases, popular mobilization promotes further unrest and conflict. Assistance in building inclusive political parties that bridge social cleavages (class, regions, ethnic groups) and are capable of leading their supporters and engaging in effective political negotiations should thus become a priority wherever political protest has played a major role in democratic transitions. Institutions that can mediate conflicts—such as supreme courts, national election commissions, or representative parliaments—are also vital factors in stemming the violent confrontation between popular groups and unpopular authorities.

VI. There is no magic bullet or golden pathway to democracy and democratic consolidation.

Finally, although it no doubt makes the job of policymakers more difficult (which they readily acknowledge), the scholars at the Stanford conference noted that there are many different paths that have led to democracy and democratic consolidation. Yet none of these are assured, as all of these paths have also failed to have the desired results. Pacts, protests, or combinations of the two, peaceful transitions and postconflict transitions, *on average* show similar rates of success in building stable democracies. Presidential and parliamentary and federal and centralized systems of government have been both successful and unsuccessful in different times and places.

The scholars noted that what matters is not so much the specific path or sequence of events leading to a transition, or the form of regime adopted, but whether the appropriate combination of factors is brought together to secure that transition, given attention to the specific context. Thus, resources should not be spent too freely in stable authoritarian contexts where change is unlikely; in postconflict states the basic conditions for progress must be secured before the transition and posttransition steps can be effective; and for countries in transition and posttransition their progress must not be neglected or starved of support in the aftermath of a transition. In addition, when opportunities arise, appropriate reactions to support change are needed in a timely fashion, and where popular mobilization is believed to be the key to change, such mobilization needs to be channeled into organizations that promote rather than undermine a peaceful and diverse civil society.

To achieve these aims it is important for democracy assistance donors

to work with local elites and democratic forces. The academic researchers expressed the view that effective democracy assistance is more a matter of facilitating than creating change, of working to encourage and maintain domestic processes, than of directing those processes.

A MULTICASE STUDY DESIGN TO GENERATE AND INVESTIGATE STRATEGIC HYPOTHESES REGARDING DEMOCRACY ASSISTANCE

For questions of strategic assessment faced by USAID—Where is spending on democracy assistance likely to pay off? How can we recognize favorable opportunities when they emerge? What kinds of obstacles are likely to prevent typical USAID democracy assistance from being fully effective? Over how long a period is assistance usefully continued under an authoritarian or semiauthoritarian regime or as a postconflict democracy seeks stability?—the committee thought that case studies could be valuable in generating and investigating hypotheses to guide USAID's allocation of DG resources.

Nonetheless, the committee was unable to agree on a firm recommendation that USAID should invest its own funds in such case studies. Since much case study research on democratization is being undertaken by academics funded by foundations and nongovernmental organizations, the committee could not reach a conclusion on how likely or unlikely this research was to be undertaken if not funded by USAID. By contrast, the improvement of its project evaluations is something that can only be done by USAID and will not be done unless the agency spends its own time and energy mandating that better evaluations be carried out. Thus the committee could agree unanimously to recommend that USAID invest in improving its project evaluations, as described in the following chapters, but not that USAID fund additional case study research of democracy assistance.

If USAID decides to invest in supporting case study research, the committee recommends using a competitive proposal solicitation process to elicit the best designs, similar to what the Strategic and Operational Research Agenda (SORA) undertook to select the design for its large-scale quantitative study (Finkel et al 2007). USAID should not specify a precise case study design but instead should specify key criteria that proposals must meet:

- **The criteria for choosing cases should be explicit and theoretically driven.**

Cases should not be selected simply because they cluster in a given region or implement a particular type of DG project. A design may focus

on a specific region or DG project, but then it should ensure that the cases within that constraint display a sufficient range of levels of USAID investment, of outcomes, and of initial contexts that they will provide a basis for identifying diverse trajectories of democratic change. The cases should be selected on criteria that will allow insights into the research question: Why did some countries make greater progress toward democracy than others, and what role did various levels of DG assistance, along with other driving factors, opportunities, and constraints, appear to play in various trajectories of progress or regress? The cases should not be selected on the arbitrary basis of a question such as: What happened in several states where USAID had DG activities?

- **The cases should include a variety of initial conditions or contexts in which USAID DG projects operate.**

The previous discussion identified three major contexts in which USAID operates programs of democracy assistance: predemocratic (authoritarian and semiauthoritarian) regimes, transition and posttransition regimes (places where authoritarian regimes no longer hold sway and democratic institutions have begun to dominate), and postconflict regimes (places where state breakdown and violence have recently occurred). Of course, postconflict regimes can be authoritarian or transitioning, and both authoritarian regimes and conflicts vary in their characteristics, as noted above. Thus this categorization only begins to frame contexts. What is crucial is that any research design acknowledge that the impact of USAID DG assistance, and prospects for democratization and stabilization, depends to a large degree on initial conditions, which vary widely across countries where USAID is asked to undertake DG projects. A good research design should not only incorporate this viewpoint but also seek to investigate how varying initial conditions affected the success of DG programming.

- **The cases should include at least one, if not several, countries in which USAID and other donors have made little or no investment in DG projects.**

Current case studies generally weigh observed outcomes in countries with DG projects against the goals of the donors. While this is sensible from one perspective—donors want to know if projects have achieved their professed goals—this is not a sound basis for gaining insights into the role that DG projects play in complex political processes. For example, a recent study of political party assistance that looked only at countries where party assistance projects were implemented concluded that such projects did little to transform political systems into more inclusive and competitive systems (Carothers 2006). Thus the donor expectations were

not met. Nonetheless, this conclusion does not allow for the possibility that party behavior might have deteriorated much more if no party assistance projects had been in place. If a comparative study that included countries with emerging political parties but few donor projects for party assistance showed that for countries without assistance, political parties tended to deteriorate more rapidly (or to more extreme levels) in regard to corruption, nepotism, factionalism, exclusion, and violence, one might argue that party assistance *is* effective, at least in holding the line against party capture by individuals or agendas adverse to democracy. The appropriate standard of comparison is thus not only what donors hoped for from DG projects but also what would have happened in the absence of such projects. By similar logic, in assessing the side effects of DG projects, including possible harm, it is important to know whether the side effects being observed are really consequences of DG assistance or are consequences that tend to arise generally as an aspect of transitions to democracy in certain contexts. Little light can be shed on this possibility unless the multicase design includes countries where DG projects were not present.

- **The cases should include countries with varied outcomes regarding democratic progress or stabilization.**

Prior USAID multicountry evaluations focus mainly on the degree to which DG projects in those countries met or fell short of donor expectations and sought to explain those shortfalls where they occurred (e.g., Carter 2001, de Zeeuw and Kumar 2006). But such evaluations did not seek out failures or the worst setbacks for detailed study. Nonetheless, sometimes the most useful information for USAID would be why projects were ineffective in particular countries. USAID has come to recognize this, but has moved too far in this direction—so that process evaluations now arise most often only when a project has failed to generate expected results. USAID needs to know both how and why DG projects succeed in various contexts and how and why they fail to generate progress in others. A rich design would include examples of both successful and unsuccessful trajectories in countries where donors have made substantial investments in DG activities.

Other Design Details

The committee does not wish to prescribe a certain number of cases for such a multicase study. Rather, that should be part of the design process and respond to the financial and time constraints chosen by USAID for the scope of the study and by the expertise and resources of the investigators. The committee does believe that a set of case studies structured by

the above criteria would provide a more comprehensive, more analytically powerful, and more valuable assessment of how democracy assistance affects countries' trajectories toward democracy than any such studies in the current literature. At the very least, it would help ensure that USAID planners have before them a diverse set of contexts and experiences from which to draw judgments, rather than the past practice of selecting five to nine cases in which USAID has intervened and then seeking to assess the results of those interventions. The committee suggests such a more structured multicase study if SORA wishes to draw on retrospective case study analysis to guide future USAID democracy programming.

However, as noted, the committee was divided over how important it would be for USAID to invest its own funds in such a research effort. Research on democracy and democracy assistance is now a rapidly growing field in the academic community (e.g., the American Political Science Association has a new section on comparative democratization), and several think tanks (e.g., Carnegie Endowment for International Peace, Center for Global Development) are supporting studies of democratization or programs to advance good governance. With the growth of interest in democracy assistance in the academic and foundation worlds, many of these issues will be investigated, and USAID may be able, in a few years, to draw on existing sets of case studies to compose a larger multicase comparison, rather than starting it from scratch. For example, a study being undertaken by the Center for Democracy, Development, and the Rule of Law at Stanford University has a design similar to that laid out in this section (CDDRL 2006:6-7).⁶ USAID may wish to simply await the completion of such academic studies over the next few years and then determine if it still needs to commission such research or if it can draw on what has already been produced in the public domain.

In sum, USAID may choose, according to its resources, to solicit proposals for comparative case studies that fulfill the above conditions, or it may choose to explore whether existing case study projects being undertaken by academics and NGOs can be tapped and combined to provide a set of case studies that meet these conditions. Either way, the committee urges USAID to encourage and examine works that go beyond the valuable, but incomplete, studies that currently focus on one or more situations in which democracy assistance has been provided. To better understand how democracy assistance affects a country's trajectory

⁶In addition to the CDDRL project, which seeks to place democracy assistance programs in the long-term and national context of diverse factors bearing on trajectories toward democracy, a number of other policy or academic works are exceptional in their breadth and quality of analysis, attending to both domestic and international factors and varying contexts and outcomes. These include particularly the work of Whitehead (1996), Carothers (1999, 2004, 2006), Mendelson and Glenn (2002), and Youngs (2004).

to democracy, it is valuable to compare trajectories with and without democracy assistance (or with relatively large and small amounts) and trajectories with varied outcomes.

However, for USAID to benefit from ongoing academic research, as well as the studies of DG assistance being undertaken by think tanks and NGOs, it will be necessary for USAID to organize regular structured interactions between such researchers and USAID DG staff. As the committee learned from the Stanford conference, academics do not always present their findings in ways that DG policy professionals find relevant; structured exchange with give and take on specific topics allows academics and professionals to bridge gaps in concepts and policy needs more effectively than passive consumption of such research. **One major service that the SORA project could perform would be to devise ways for the more regular introduction of scholars' research on democracy into structured discussions with USAID DG personnel.**

Besides such a multicountry case study design, the committee also believes that there are other ways for USAID to learn from its past DG activities. These include discussions of outside studies of DG assistance, such as those undertaken by the Carnegie Endowment (e.g., Carothers 2006, 2007) or other nations' development agencies, statistical analyses of international data, and surveys. These also include making better use of the experience of USAID DG mission personnel by engaging in regular meetings in which DG officers could share and discuss their own experiences with democracy assistance. Although not adequate for determining the impact of specific projects, such sources can provide valuable insights regarding problems of program implementation, responses to rapidly changing conditions in the field, issues in the reception of DG programs, or the shifting contexts in which such programs are carried out. **USAID should include these varied sources of information as part of the regular organizational learning activities recommended in Chapter 8.**

CONCLUSIONS

The committee found that much can be gleaned from existing case studies of democracy and governance. These studies of particular programs, or of DG assistance in specific regions, shed light on how DG programs have operated in various settings and whether they met the expectations of donors or participants. Yet for all their strengths it is often difficult to solve the problem of causal attribution of specific outcomes to DG activities with this type of research. This is particularly true of studies that attempt to discern the causal impact of a particular project or set of projects on democracy by focusing only on the unfolding of DG projects within a single country or across a set of countries.

The committee thus recommends the use of more diverse and theoretically structured clusters of case studies of democratization and democracy assistance to develop hypotheses to guide democracy assistance planning in a diverse range of settings. Whether USAID chooses to support such studies or gather them from ongoing academic research, it is important to look at how democracy assistance functions in a range of different initial conditions and trajectories of political change. Such case studies should seek to map out long-term trajectories of political change and to place democracy assistance in the context of national and international factors affecting those trajectories, rather than focus mainly on specific democracy assistance programs.

REFERENCES

- Abbink, J., and Hesselting, G., eds. 2000. *Election Observation and Democratization in Africa*. New York: St. Martin's Press.
- Asia Watch. 2002. Cambodia's Commune Elections: Setting the Stage for the 2003 National Elections. *HRW Index* 14(4). Available at: [http://www.hrw.org/reports/2002/camb\)402/](http://www.hrw.org/reports/2002/camb)402/). Accessed on January 10, 2008.
- Barro, R.J. 1999. Determinants of Democracy. *Journal of Political Economy* 6:158-183.
- Barya, J.J., Opolot, S.J., and Otim, P.O. 2004. The Limits of "No Party" Politics: The Role of International Assistance in Uganda's Democratisation Process. Working Paper 28. Conflict Research Unit, Netherlands Institute of International Relations, Clingendael.
- Berman, S.E. 1997. Civil Society and the Collapse of the Weimar Republic. *World Politics* 49(3):401-429.
- Berman, S.E. 2001. Modernization in Historical Perspective: The Case of Imperial Germany. *World Politics* 53(2):431-462.
- Blair, H., and Hansen, G. 1994. *Weighing in on the Scales of Justice: Strategic Approaches for Donor-Supported Rule of Law Programs*. CDIE Program and Operations Assessment Report No. 7. Washington, DC: USAID.
- Boix, C., and Stokes, S. 2003. Endogenous Democratization. *World Politics* 55(4):517-549.
- Carothers, T. 1999. *Aiding Democracy Abroad: The Learning Curve*. Washington, DC: Carnegie Endowment.
- Carothers, T. 2004. *Critical Mission: Essays on Democracy Promotion*. Washington, DC: Carnegie Endowment.
- Carothers, T. 2006. *Confronting the Weakest Link: Aiding Political Parties in New Democracies*. Washington, DC: Carnegie Endowment.
- Carothers, T. 2007. The "Sequencing" Fallacy. *Journal of Democracy* 18(1):12-27.
- Carter, L. 2001. Linking USAID Democracy Program Impact to Political Change: A Synthesis of Findings from Three Case Studies. Revised draft (unpublished).
- Carter, L., Silver, R., and Smith, Z. 2003. *Linking USAID Democracy Program Impact to Political Change: A Synthesis of Findings from Six Case Studies*. Washington, DC: Management Systems International.
- CDDRL (Center for Democracy, Development, and the Rule of Law). 2006. Project Prospectus. Unpublished.
- Dobbins, J. 2003. *America's Role in Nation Building: From Germany to Iraq*. Santa Monica, CA: RAND.

- Dougherty, B.K. 2004. Searching for Answers: Sierra Leone's Truth & Reconciliation Commission. *African Studies Quarterly* 8(1). Online. Available at <http://web.africa.ufl.edu/asq/v8/v8ia3.htm>.
- Doyle, A.C. 1998. *The Adventure of the Beryl Coronet, The Adventures of Sherlock Holmes*. Oxford World's Classics. New York: Oxford University Press.
- Epstein, D., Bates, R., Goldstone, J.A., Kristensen, I., and Halloran, S. 2006. Democratic Transitions. *American Journal of Political Science* 50:551-569.
- Finkel, S.E., Pérez-Liñán, A., and Seligson, M.A. 2007. The Effects of U.S. Foreign Assistance on Democracy Building, 1990-2003. *World Politics* 59(3):404-439.
- George, A., and Bennett, A. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.
- Goldstone, J.A. 1998. Initial Conditions, General Laws, Path-Dependence, and Explanation in Historical Sociology. *American Journal of Sociology* 104:829-845.
- Goldstone, J.A. 2003. Comparative Historical Analysis and Knowledge Accumulation in the Study of Revolutions. Pp. 41-90 in *Comparative Historical Analysis*, D. Reuschemeyer and J. Mahoney, eds. Cambridge: Cambridge University Press.
- Goldstone, J.A. 2007. Trajectories of Democracy and Development: New Insights from Graphic Analysis. Paper presented to Wilton House, UK, October 23.
- Goldstone, J.A., and Ulfelder, J. 2004. How to Construct Stable Democracies. *Washington Quarterly* 28(1):9-20.
- Haggard, S., and Kauffman, R.R. 1995. *The Political Economy of Democratic Transitions*. Princeton, NJ: Princeton University Press.
- Kumar, K. 1998. Post-Conflict Elections, Democratization, and International Assistance. Boulder: Lynne Rienner.
- Lippman, H., and Emmert, J. 1997. *Assisting Legislatures in Developing Countries: A Framework for Program Planning and Implementation*. Washington, DC: USAID.
- McFaul, M. 2005. Transitions from Post-Communism. *Journal of Democracy* 16(3):5-19.
- McFaul, M. 2006. *The 2004 Presidential Elections in Ukraine and the Orange Revolution: The Role of U.S. Assistance*. Washington, DC: USAID, Office for Democracy and Governance.
- McMahon, E., Beale, S., and Menelik-Swanson, G. 2004. *Ethiopia Pre-Election Assessment Report*. Washington, DC: International Foundation for Election Systems. Available at: <http://www.ifes.org/publication/f6f42ace604bfb37be74675f7d4d002b/Ethiopia.pdf>.
- Mendelson, S.E., and Glenn, J.K., eds. 2002. *The Power and Limits of NGOs: A Critical Look at Building Democracies in Eastern Europe and Eurasia*. New York: Colombia University Press.
- O'Neill, W.G. 2003. International Human Rights Assistance: A Review of Donor Activities and Lessons Learned. Working Paper No. 18. The Hague, Netherlands: Clingendael Institute.
- Przeworski, A., Alvarez, M.E., Cheibub, J.A., and Limongi, F. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950-1990*. Cambridge: Cambridge University Press.
- Rice, C. 2006. Transformational Diplomacy. U.S. Department of State. Available at: <http://www.state.gov/r/pa/prs/ps/2006/59339.htm>.
- Whitehead, L. ed. 1996. *The International Dimensions of Democracy: Europe and the Americas*. Oxford: Oxford University Press.
- Youngs, R. 2004. *International Democracy and the West: The Role of Governments, NGOs and Multinationals*. Oxford: Oxford University Press.
- Zakaria, F. 2004. *The Future of Freedom*. New York: Norton.
- de Zeeuw, J., and Kumar, K. 2006. *Promoting Democracy in Post-Conflict Societies*. Boulder: Lynne Rienner.

Methodologies of Impact Evaluation

INTRODUCTION

This chapter presents a guide to impact evaluations as they are currently practiced in the field of foreign assistance. The committee recognizes, as stated before, that the application of impact evaluations to foreign assistance in general, and to democracy and governance (DG) projects in particular, is controversial. The purpose of this chapter is thus to present the range of impact evaluation designs, as a prelude to the results of the committee's field teams' exploration of their potential application as part of the mix of evaluations and assessments undertaken by the U.S. Agency for International Development (USAID) presented in the next two chapters.

The highest standard of credible inference in impact evaluation is achieved when the number of people, villages, neighborhoods, or other groupings is large enough, and the project design flexible enough, to allow randomized assignment to treatment and nontreatment groups. Yet the committee realizes that this method is often not practical for many DG projects. Thus this chapter also examines credible inference designs for cases where randomization is not possible and for projects with a small number of units—or even a single case—involved in the project.

Some of the material in this chapter is somewhat technical, but this is necessary for this chapter to serve, as the committee hopes it will, as a guide to the design of useful and credible impact evaluations for DG missions and implementers. The technical terms used here are defined in the chapter text and also in the Glossary at the end of the report. Also,

examples are provided to show how such designs have already been implemented in the field for various foreign assistance and democracy assistance programs.

IMPORTANCE OF SOUND AND CREDIBLE IMPACT EVALUATIONS FOR DG ASSISTANCE

As discussed in some detail in Chapter 2, until 1995 USAID required evaluations of all its projects, including those in DG, to assess their effectiveness in meeting program goals. Most of the evaluations, however, were process evaluations: post-hoc assessments by teams of outside experts who sought to examine how a project unfolded and whether (and why) it met anticipated goals. While these were valuable examinations of how projects were implemented and their perceived effects, such evaluations generally could not provide the evidence of impact that would result from sound impact assessments. This was because in most cases they lacked necessary baseline data from before the project was begun and because in almost all cases they did not examine appropriate comparison groups to determine what most likely would have occurred in the absence of the projects (see Bollen et al [2005] for a review of past DG evaluations).

As noted, the number of such evaluations undertaken by USAID has declined in recent years. Evaluations are now optional and are conducted mainly at the discretion of individual missions for specific purposes, such as when a major project is ending and a follow-on is expected or when a DG officer feels that something has “gone wrong” and wants to understand and document the reasons for the problem. Such special evaluations can have substantial value for management purposes, but the committee believes that USAID is overlooking a major opportunity to learn systematically from its experience about project success and failure by not making impact evaluations a significant part of its monitoring and evaluation (M&E) activities where appropriate and feasible. Such impact evaluations could be particularly useful to provide insights into the effects of its largest-scale and most frequently used projects and to test key development hypotheses that guide its programming.

There are three fundamental elements of sound and credible impact evaluations. First, such evaluations require measures relevant to desired project *outcomes*, not merely of project activity or outputs. Second, they require good *baseline*, in-process, and endpoint measures of those outcomes to track the effects of interventions over time. Finally, they require *comparison* of those who receive assistance with appropriate nontreatment groups to determine whether any observed changes in outcomes are, in fact, due to the intervention.

The committee’s discussions with USAID staff, contractors for USAID,

and our own field study of USAID missions have shown that, even within the current structure of project monitoring, USAID is already engaged in pursuing the first and second requirements. While in some cases progress remains to be made on devising appropriate outcome measures and in ensuring the allocation of time and resources to collect baseline data, USAID has generally recognized the importance of these tasks. These efforts do vary from mission to mission, according to their available resources and priorities, so considerable variation remains among missions and projects in these regards.

However, the committee found that there is little or no evidence in current or past USAID evaluation practices that indicates the agency is making regular efforts to meet the third requirement—comparisons. With rare exceptions, USAID evaluations and missions generally do not allocate resources to baseline and follow-up measurements on nonintervention groups. Virtually all of the USAID evaluations of which the committee is aware focus on studies of groups that received USAID DG assistance, and estimates of what would have happened in the absence of such interventions are based on assumptions and subjective judgments, rather than explicit comparisons with groups that did not receive DG assistance. It is this almost total absence of comparisons with nontreated groups, more than any other single factor, that should be addressed in order to draw more credible and powerful conclusions about the impact of USAID DG projects in the future.

To briefly illustrate the importance of conducting baseline and follow-up measurements for both treated and nontreated comparison groups, consider the following two simple examples:

1. A consulting firm claims to have a training program that will make legislators more effective. To demonstrate the program's effectiveness, the firm recruits a dozen legislators and gives them all a year of training. The firm then measures the number of bills those legislators have introduced in parliament in the year prior to the training and the number of bills introduced in the year following the training and finds that each legislator increased the number of bills he or she had introduced by 30 to 100 percent! Based on this the consultants claim they have demonstrated the efficacy of the program.

Yet to know whether or not the training really was effective, we would need to know how much each legislator's performance would have changed if he or she had *not* taken the training program. One way of answering this question is to compare the performance of the legislators who were trained to the performance of a comparable set of legislators who were not. When someone points this out to the consultants and they go back and measure the legislative activity of *all* the legislators for

the prior year, they find that the legislators who were *not* in the training group introduced, on average, exactly the same number of bills as those who were trained.

What has happened? It is possible that the increase in the number of bills presented by all legislators resulted from greater experience in office, so that everyone introduces more bills in his or her third year in office than in the first year. Or there may have been a rule change, or policy pressures, that resulted in a general increase in legislative activity. Thus it is entirely possible that the observed increase in legislative activity by those trained had nothing to do with the training program at all, and the program's effect might have been zero.

Or it is possible that those legislators who signed up for the program were an unusual group. They might have been those legislators who were already the most active and who wanted to increase their skills. Thus the program might have worked for them but would not have worked for others. Another possibility is that the legislators who signed up were those who were the *least active* and who wanted the training to enable them to "catch up" with their more active colleagues. In this case the results *do* show merit to the training program, but again it is not clear how much such a program would help the average legislator improve.

The only way to resolve these various possibilities would be to have taken measures of legislative activity before and after the training program for both those legislators in the program and those not in the program. While it would be most desirable to have randomly assigned legislators to take the training or not, that is not necessary for the before and after comparison measures to still yield valuable and credible information. For example, even if legislators themselves chose who would receive the training, we would want to know whether the trained group had previously been more active, or less active, than their colleagues not receiving training. We could also then make statistical adjustments to the comparison, reflecting differences in prior legislative activity and experience between those who were trained and those who were not, to help determine what the true impact of the training program was, net of other factors that the training could not affect.

In short, simply knowing that a training program increased the legislative activity of those trained does not allow one to choose between many different hypotheses regarding the true impact of that program, which could be zero or highly effective in providing "catch-up" skills to legislators who need them. The only way to obtain sound and credible judgments of a program's effect is with before and after measurements on both the treatment and the relevant nontreatment groups.

2. The same consulting firm also claims to have a program that will increase integrity and reduce corruption among judges. To test the

program's effectiveness, the firm recruits a dozen judges to receive the program's training for a year. When the consultants examine the rate of perceived bribery and corruption, or count cases thrown out or settled in favor of the higher status plaintiff or defendant, in those courts where the judges were trained, they find that there has been no reduction in those measures of corruption. On this basis the donor might decide that the program did not work. However, to really reach this conclusion, the donor would have to know whether, and how much, corruption would have changed if those judges had not received the training. When the donor asks for data on perceived bribery and corruption, or counts of cases thrown out or settled in favor of higher status plaintiffs or defendants, in other courts it turns out to be much higher than in the courts where judges did receive the training.

Again, the new information forces us to ask: What really happened? It is possible that opportunities for corruption increased in the country, so that most judges moved to higher levels of corruption. In this case the constant level of corruption observed in the courts whose judges received training indicated a substantially greater ability to resist those opportunities. So, when properly evaluated against a comparison group, it turns out that the program *was, in fact, effective*. To be sure, however, it would be valuable to also have baseline data on corruption levels in the courts whose judges were not trained; this would confirm the belief that corruption levels increased generally except in those courts whose judges received the program. Without such data it is not known for certain whether this is true or whether the judges who signed up for the training were already those who were struggling against corruption and who started with much lower rates of corruption than other courts.

These examples underscore the vital importance of *comparisons with groups not receiving the treatment* in order to avoid misleading errors and to accurately evaluate project impacts. From a public policy standpoint, the cost of such errors can be high. In the examples given here, it might have caused aid programs to waste money on training programs that were, in fact, ineffective. Or it might have led to cuts in funding for anticorruption programs that were, in fact, highly valuable in preventing substantial increases in corruption.

This chapter discusses how best to obtain comparisons for evaluating USAID democracy assistance projects. Such comparisons range from the most rigorous possible—comparing randomly chosen treatment and nontreatment groups—to a variety of less exacting but still highly useful comparisons, including multiple and single cases, time series, and matched case designs. It bears repeating: The goal in all of these designs is to evaluate projects by using appropriate comparisons in order to increase confidence in drawing conclusions about cause and effect.

PLAN OF THIS CHAPTER

The chapter begins with a discussion of what methodologists term “internal” and “external” validity. Internal validity is defined as “the approximate truth of inferences regarding cause-effect or causal relationships” (Trochim and Donnelly 2007:G4). The greater the internal validity, the greater the confidence one can have in the conclusions that a given project evaluation reaches. The paramount goal of evaluation design is to maximize internal validity. External validity refers to whether the conclusions of a given evaluation are likely to be applicable to other projects and thereby contribute to understanding in a general sense what works and what does not. Given that USAID implements similar projects in multiple country settings, the external validity of the findings of a given project evaluation is particularly important. This section of the chapter also stresses the importance of what the committee terms “building knowledge.”

The second part of the chapter outlines a typology of evaluation methodologies that USAID missions might apply in various circumstances to maximize their ability to assess the efficacy of their programming in the DG area. Large N randomized designs permit the most credible inferences about whether a project worked or not (i.e., the greatest internal validity). By comparison, the post-hoc assessments that are the basis of many current and past USAID evaluations provide perhaps the least reliable basis for inferences about the actual causal impact of DG assistance. Between these two ends of the spectrum lie a number of different evaluation designs that offer increasing levels of confidence in the inferences one can make.

In describing these various evaluation options, the approach taken in this chapter is largely theoretical and academic. Evaluation strategies are compared and contrasted based on their methodological strengths and weaknesses, not their feasibility in the field. While a first step is taken at the end of the chapter in the direction of exploring whether the most rigorous evaluation design—large N randomized evaluation—is feasible for many DG projects, a more extensive treatment of this key question is reserved for the chapters that follow, when the committee presents the findings of its field studies, in which the feasibility of various impact evaluation designs is explored for current USAID DG programs with mission directors and DG staff.

POINTS OF CLARIFICATION

Before plunging into the discussion of evaluation methodologies, a few important points of clarification are needed. First, it should be clear that the committee’s focus on impact evaluations is not intended to deny

the need for, or imply the unimportance of, other types of M&E activities. The committee recognizes that monitoring is vital to ensure proper use of funds and that process evaluations are important management tools for investigating the implementation and reception of DG projects. This report focuses on how to develop impact evaluations because the committee believes that at present this is the most underutilized approach in DG program evaluations and that therefore USAID has the most to gain if it is feasible to add sound and credible impact evaluations to its portfolio of M&E activities.

Second, the committee recognizes that not all projects need be, or should be, chosen for the most rigorous forms of impact evaluation. Doing so would likely impose an unacceptably high cost on USAID's DG programming. The committee is therefore recommending that such evaluations initially only be undertaken for a select few of USAID's DG programs, a recommendation emphasized in Chapter 9. The committee does believe, however, that DG officers should be aware of the potential value of obtaining baseline and comparison group information for projects to which they attach great importance, so that they can better decide how to develop the mix of M&E efforts across the various projects that they oversee.

Third, before beginning the task of evaluating a project, precisely what is to be evaluated must be defined. Evaluating a project requires the identification of the specific intervention and a set of relevant and measurable outcomes thought to result from that policy intervention. Even this apparently simple task can pose challenges, since most DG programs are complex (compound) interventions, often combining several activities (e.g., advice, formal training, monetary incentives) and are often expected to produce several desired outcomes. A project focused on the judiciary, for example, may include a range of different activities intended to bolster the independence and efficiency of the judiciary in a country and might be expected to produce a variety of outcomes, including swifter processing of cases, greater impartiality among plaintiffs and defendants, greater conformity to statutes or precedents, and greater independence vis-à-vis the executive. The evaluator must therefore decide whether to test the whole project or parts of the project or whether it would make sense, as discussed further below, to reconfigure the project to allow for clearer impact evaluation of specific interventions.

As USAID's primary focus will always be on program implementation, rather than evaluation per se, evaluators will need to respond to the challenges posed by often ambitious and multitasked programs.

At this point, a note on terminology is required. As noted above, an "activity" is defined as the most basic sort of action taken in the field, such as a training camp, a conference, advice rendered, money tendered,

and so forth. A “project” is understood to be an aggregation of activities, including all those mentioned in specific USAID contracts with implementers, such as in requests for proposals and in subsequent documents produced in connection with these projects. A project can also be referred to as an “intervention” or “treatment.”

The question of what constitutes an appropriate intervention is a critical issue faced by all attempts at project evaluation. A number of factors impinge on this decision.

Lumping activities within a given project together for evaluation often makes sense. If all parts of a program are expected to contribute to common outcomes, and especially if the bundled activities will have a stronger and more readily observed outcome than the separate parts, then treating the set of activities together as a single intervention may be the best way to proceed.

In other cases, trying to separate various activities and measuring their impact may be preferred. The value of disaggregation seems clear from the standpoint of impact evaluation. After all, if only one part of a five-part program is in fact producing 90 percent of the observed results, this would be good to know, so that only that one part continues to be supported. But whether or not such a separation seems worth testing really depends on whether it is viable to offer certain parts of a project and not others. Sometimes it is possible to test both aggregated and disaggregated components of a project in a single research design. This requires a sufficient number of cases to allow for multiple treatment groups. For example, Group A could receive one part of a program, Group B could receive two parts of a program, Group C could receive three parts of a program, and another group would be required as a control. In this example, three discrete interventions and their combination could be evaluated simultaneously.

Many additional factors may impinge on the crafting of an appropriate design for impact evaluation of a particular intervention. These are reviewed in detail in the subsequent section. The committee understands that there is no magic formula for deciding when an impact evaluation might be desirable or which design is the best trade-off in terms of costs, need for information, and policy demands. What is clear, however, is that since impact evaluations are, in effect, tests of the hypothesis that a given intervention will create different outcomes than would be observed in the absence of that intervention, how well one specifies that hypothesis greatly influences what one will find at the end of the day. The question asked determines the sort of answers that can be received. The committee wants to flag this as a critical issue for USAID policymakers and project implementers to consider; further suggestions are given in Chapters 8 and 9 for how this could be addressed as part of an overall Strategic and

Operational Research Agenda project for learning about DG program effectiveness to guide policy programming.

INTERNAL VALIDITY, EXTERNAL VALIDITY, AND BUILDING KNOWLEDGE

Internal Validity

A sound and credible impact evaluation has one primary goal: to determine the impact of a particular project in a particular place at a particular time. This is usually understood as a question of internal validity. In a given instance, what causal effect did a specific policy intervention, X, have on a specific outcome, Y? This question may be rephrased as: If X were removed or altered, would Y have changed?

Note that the only way to answer this question with complete certainty is to go back in time to replay history without the project (called the “the counterfactual”). Since that cannot be done, we try to come as close as possible to the “time machine” by holding constant any background features that might affect Y (the *ceteris paribus* conditions) while altering X, the intervention of interest. We thus replay the scenario under slightly different circumstances, observing the result (Y).

It is in determining how best to simulate this counterfactual situation of replaying history without the intervention that the craft of evaluation design comes into play. Indeed, a large literature within the social sciences is devoted to this question—often characterized as a question of causal assessment or research design (e.g., Shadish et al 2002, Bloom 2005, Duflo et al 2006b). The following section attempts to reduce this complicated set of issues down to a few key ingredients, recognizing that many issues can be treated only superficially.

Consider that certain persistent features of research design may assist us in reaching conclusions about whether X really did cause Y: (1) interventions that are simple, strong, discrete, and measurable; (2) outcomes that are measurable, precise, determinate, immediate, and multiple; (3) a large sample of cases; (4) spatial equivalence between treatment and control groups; and (5) temporal equivalence between pre- and posttests. Each of these is discussed in turn.

1. The intervention: discrete, with immediate causal effects, measurable. A discrete intervention that registers immediate causal effects is easier to test because only one pre- and posttest is necessary (perhaps only a posttest if there is a control group and trends are stable or easily neutralized by the control). That is, information about the desired outcome is collected before and after the intervention. By contrast, an intervention

that takes place gradually, or has only long-term effects, is more difficult to test. A measurable intervention is, of course, easier to test than one that is resistant to operationalization (i.e., must be studied through proxies or impressionistic qualitative analysis).

2. The outcome(s): measurable, precise, determinate, and multiple.

The best research designs feature outcomes that are easily observed, that can be readily measured, where the predictions of the hypotheses guiding the intervention are precise and determinate (rather than ambiguous), and where there are multiple outcomes that the theory predicts, some of which may pertain to causal processes rather than final policy outcomes. The latter is important because it provides researchers with further evidence by which to test (confirm or disconfirm) the underlying hypothesis linking the intervention to the outcome and to elucidate its causal mechanisms.

3. Large sample size. *N* refers here to the number of cases that are available for study in a given setting (i.e., the sample size). A larger *N* means that one can glean more accurate knowledge about the effectiveness of the intervention, all other things being equal. Of course, the cases within the sample must be similar enough to one another to be compared; that is, the posited causal relationship must exist in roughly the same form for all cases in the sample or any dissimilarities must be amenable to post-hoc modeling. Among the questions to be addressed are: How large is the *N*? How similar are the units (cases) in respects that might affect the posited causal relationships? If dissimilar, can these heterogeneous elements be neutralized by some feature of the research design (see below)?

4. Spatial equivalence (between treatment and control groups). By pure spatial comparisons what is meant are controls that mirror the treatment group in all ways that might affect the posited causal relationship. The easiest way to achieve equivalence between these two groups is to choose cases randomly from the population. Sometimes, nonrandomized selection procedures can be achieved, or exist naturally, that provide equivalence, but this is relatively rare. The key question to ask is always: How similar are the treatment and control groups in ways that might affect the intended outcome? This is often referred to as “pretreatment equivalence.” Other important questions include: Can the treatment cases be chosen randomly, or through some process that approximates random selection? Can the equivalence initially present at the point of intervention between treatment and control groups be maintained over the life of the study (i.e., over whatever time is relevant to observe the putative causal effects)? This may be referred to as “posttreatment equivalence.”

5. Temporal equivalence (between pre- and posttests). Causal attribution works by comparing spatially and/or temporally. This is usually done through pre- and posttreatment tests (i.e., measurements of the outcome before and after the intervention, creating two groups, the pre-

intervention group and the postintervention group. Of course, it is the same case, or set of cases, observed at two points in time. However, such comparisons (in isolation from spatial controls) are useful only when the case(s) are equivalent in all respects that might affect the outcome (except, of course, insofar as the treatment itself). More specifically, this means that (1) the effects of the intervention on the case(s) are not obscured by confounders, which are other factors occurring at roughly the same time as the intervention which might affect the outcome, and (2) the outcome under investigation either is stable or has a stable trend (so that the effect of the intervention, if any, can be observed). Note that when there is a good spatial control these issues are less important. By contrast, when there is *no* spatial control, they become absolutely essential to the task of causal attribution. For temporal control the key questions to ask are: Are comparable pre- and posttests possible? Is it possible to collect data for a longer period of time so that, rather than just two data points, one can construct a longer time series? Are there trends in the outcome that must be taken into account? If trends are present, are they fairly stable? Can we anticipate that this stability will be retained over the course of the research (in the absence of any intervention)? Is the intervention correlated (temporally) with other changes that might obscure causal attribution?

External Validity

External validity is the generalizability of the project beyond a single case. To provide policymakers at USAID with relevant information, the results of a project evaluation should be generalizable; that is, they must be true (or plausibly true) beyond the case under study. Recall that we understand that impact *evaluation* (as opposed to project *monitoring*) will most likely be an occasional event applied to a set of the most important and most frequently used projects, not one routinely undertaken for all projects. This means that the value of the evaluation is to be found in the guidance it may offer policymakers in designing projects and allocating funds over the long term and across the whole spectrum of countries in which USAID works.

There will always be questions about how much one can generalize about the impact of a project. The fact that a project worked in one place, at one time, may or may not indicate its possible success in other places and at other times. The committee recognizes that the design of USAID projects and the allocation of funds are a learning process and the political situation and opportunities for intervention in any given country are a moving target. Even so, project officers must build on what they know, and this knowledge is largely based on the experiences of projects that are currently in operation around the world. Some projects are perceived

to work well while others are perceived to work poorly or not at all. It is these general perceptions of “workability” that are the concern here. With a number of sound impact evaluations of a specific type of project in several different settings, USAID would be able to learn more from its interventions, rather than rely solely on the experiences of individuals.

To maximize the utility of such impact evaluations, each aspect of the research design must be carefully considered. Two factors are paramount: *realism* in evaluation design and careful *case selection*.

Realism means that the evaluation of a project should conform as closely as possible to existing realities on the ground; otherwise, it is likely to be dismissed as an exercise with little utility for USAID officers in the field. “Realities” refers to the political facts at home and abroad, the structure of USAID programming, and any other contextual features that might be encountered when a project is put into operation. The committee recognizes that some factors on the ground may need to be altered in order to enhance the internal validity of a research design, a matter addressed below. Yet for purposes of external validity in the policymaking world of USAID, these factors should be kept to a minimum.

Case selection refers to how cases—activities or interventions—are chosen for evaluation. Several strategies are available, each with a slightly different purpose. However, all relate to the achievement of external validity.

The most obvious strategy is to choose a *typical* case, a context that is, so far as one can tell, typical of that project’s usual implementation and also one that embodies a typical instance of posited cause-and-effect relationships. Otherwise, it may be difficult to generalize from that project’s experience.

A second strategy is known as the *least likely* (or most difficult) case. If one is fairly confident of a project’s effectiveness, perhaps because other studies have already been conducted on that subject, confidence can be enhanced by choosing a case that would not ordinarily be considered a strong candidate for project success. If the project is successful there, it is likely to be successful anywhere (i.e., in “easier” circumstances). Alternatively, if the project fails in a least-likely setting, then one has established a boundary for the population of cases to which the project may plausibly apply.

A third strategy is known as the *most likely* case. As implied, this kind of case is the inverse of the previous: It is one where a given intervention is believed most likely to succeed. This kind of case is generally useful only when the intervention, against all odds, is shown by a careful impact evaluation to have little or no effect (otherwise, common wisdom is confirmed). Failure in this setting may be devastating to the received wisdom, for it would have shown that even when conditions are favorable the project still does not attain its expected result.

Other strategies of case selection are available; further strategies and a more extended discussion can be found in Chapter 5 of Gerring (2007). For the purposes of project evaluation at USAID, however, these three appear likely to be the most useful.

Because of the varied contexts in which even “typical” USAID projects are implemented, it would be best to conduct impact evaluations to determine the effects of such projects in several different places. Ideally, USAID could choose a “typical” case, a least likely case, and a most likely case for evaluation to determine whether a project is having its desired impact. Even if this spread is not readily available, choosing two or three different sites to evaluate widely used projects would help address concerns about generalizability more effectively than using only a single site for an impact evaluation.

Building Knowledge

It is important to keep in mind that no single evaluation is likely to be regarded as complete evidence for or against a project, nor should it. Regardless of how carefully an evaluation is designed, there is always the possibility of random error—factors at work in a country or some sector of a country that cannot be controlled by carefully constructed evaluation designs. More importantly, there is always the possibility that an intervention may work differently in one setting than it does in others. Thus the process of evaluating projects should always involve multiple evaluations of the same basic intervention. This means that strategies of evaluation must take into account the body of extant knowledge on a subject and the knowledge that may arise from future studies (supported by USAID, other agencies, or the academic community). This is the process of building knowledge. The most successful companies in the private sphere tend to be “learning organizations” that constantly build knowledge about their own activities (Senge 2006). This process may be disaggregated into four generic goals: building *internal validity*, building *external validity*, building *better project design*, and building *new knowledge*.

The first three issues may be understood as various approaches to “replication.” If USAID is concerned about the internal validity of an impact evaluation, subsequent evaluations should replicate the original research design as closely as possible. If USAID is concerned about the external validity of an evaluation, then replications should take place in different sites. If USAID is concerned with the specific features of a project, replications should alter those features while keeping other factors constant. The fourth issue departs from the goal of replication; here the goal is to unearth new insights into the process of development and the degree to which it may be affected by USAID policies. In this instance it is no longer so important to replicate features of previous evaluations.

Even so, the committee emphasizes that the important features of a research design—the treatment, the outcomes anticipated to result from the treatment, and the setting—should be standardized as much as possible across each evaluation. Doing so helps ensure that the results of the evaluation will be comparable to evaluations of similar projects, so that knowledge accumulates about that subject. If the treatments and evaluation designs change too much from evaluation to evaluation, less can be learned.

Using impact evaluations in no way reduces the need for sound judgment from experienced DG staff; detailed knowledge of the country and specific conditions is essential for creating a good impact design. More generally, there are often external events that can have consequences for an ongoing project or its evaluation. In such cases an experienced DG officer will need to appraise the effect of these events on the project's process and outcomes. However, an appropriate mix of evaluations offers better information about projects on which DG staff can create new, more effective policy.

A TYPOLOGY OF IMPACT EVALUATION DESIGNS

A major goal of this chapter is to identify a reasonably comprehensive, yet also concise, typology of research designs that might be used to test the causal impact of projects supported by USAID's DG office. Six basic research designs seem potentially applicable: (1) large N with random assignment of the project¹; (2) large N comparison without randomized assignment of the project; (3) small N with randomized assignment of the project; (4) small N without randomized assignment of the project; (5) $N = 1$, where USAID has control of where or when the project is put in place; and (6) $N = 1$, where USAID has little control over where or when the project is placed. Each option is summarized in Table 5-1.

Each research design shown in the table shares a dedicated effort to collect pre- and posttreatment measures of the policy outcomes of interest. Hitherto, baseline measurements have been an inconsistent part of USAID evaluations (Bollen et al 2005); although baseline data are generally supposed to be collected as part of current program monitoring, the quality may vary substantially. The absence of good baseline data makes it much more difficult to demonstrate a causal effect. No project can be adequately tested without a good measurement of the outcome of interest prior to the policy intervention. Naturally, such a measurement should be paired with a corresponding measurement of the outcome after the policy interven-

¹Randomized assignment of a treatment is often called an experiment in texts on research design (see, e.g., Trochim and Donnelly 2007).

TABLE 5-1 A Typology of Suggested Research Designs

Available Units (N)	Manipulability	Pre-/Posttests	Suggested Research Design
Large	Yes	Yes	Large N randomization
Large	No	Yes	Large N comparison
Small	Yes	Yes	Small N randomization
Small	No	Yes	Small N comparison
1	Yes	Yes	N = 1 study (manipulable)
1	No	Yes	N = 1 comparison

tion. (See Chapters 6 and 7 for further discussion of appropriate measures of outcomes, with examples from the committee’s field visits.) Together, these provide pre- and posttests of the policy intervention.

In the large N randomized assignment design—but *only* in that case—it is possible to evaluate project outcomes even in the absence of baseline data, as shown, for example, in Hyde (2006), where she evaluated the impact of election monitors from observed differences in the votes received by opposition parties in precincts with and without the randomly assigned monitors. However, this procedure always *assumes* that the intervention and control groups would show similar outcomes in the absence of any intervention. It is better, wherever possible, to check this assumption with baseline data. This is particularly important when the number of cases is modest and full randomization is not possible, and many other factors besides the intervention can affect outcomes. Even in the case of the large N randomization, baseline data are often useful for checking the assumptions on which programming is based, or for planning or evaluating other projects later.²

The six research design options are *distinguishable* from one another along two key dimensions: (1) the number of units (N) available for analysis and (2) USAID’s capacity to manipulate key features of the project’s design and implementation. Usually, the capacity to evaluate projects is enhanced when N is large (i.e., when there are a large number of individuals, organizations, or governments that can be compared to one another) and when the project can be implemented in a randomized way. The large N randomized intervention is thus regarded as the “gold standard” of project evaluation methods (Wholey et al 2004). Each step away from the large N randomized design generally involves a loss in inferential power or, in other words, less confidence in the ability to make inferences about causal impact based on the results of the evaluation.

Even so, this certainly does not imply, and the committee is not argu-

²For examples, see the research papers on the Poverty Action Lab of MIT webpage: <http://www.povertyactionlab.com/papers/>.

ing, that the large N randomized intervention is the *only* viable evaluation tool available to USAID.³ If this were the case, many projects—and the millions of dollars used to fund them—could not be the subject of impact evaluations. It is for this reason that the committee offers a longer list of options than is recognized by many current texts on project evaluation (e.g., Bloom 2005, Duflo et al 2006b). But the results of the committee's visits to USAID offices in the field, review of USAID documents, and discussions with USAID DG officials and implementers suggest that using randomization is feasible at least in theory in many instances, which would greatly enhance the ability to evaluate the impacts of a project.

Of course, no simple classification of types can hope to address all the research design issues raised by the multifaceted programs supported by USAID's DG portfolio of projects. Arguably, every policy intervention is in some respects unique and thus poses different research design issues. Measuring impact is not easy. The committee offers the foregoing typology as a point of departure, a set of categories that capture the most salient features of different policies now supported by the USAID DG office, and the ways in which the causal impact of these policies might be feasibly evaluated. Citations in the text to existing work on these subjects should provide further guidance for project officers and implementers, although the literature on large N randomized treatment research designs is much more developed than the literature on other subjects.

1. Large N Randomized Evaluation

The ideal research design is the randomized impact evaluation. Because of its technical demands, this approach should be employed where USAID DG officials have a strong interest in finding out the impact of an important project, especially those that are implemented in a reasonably similar form across countries (e.g., decentralization initiatives, civic education projects, election monitoring efforts). Here, a large number of units are divided by *random selection* into treatment and control groups, a treatment is administered, and any differences in outcomes across the two groups are examined for their significance. Randomizing the treatment attempts to break a pool of possible treated units into two groups that are similar, indeed indistinguishable, before the treatment. Then, after the treatment, measurement on the desired outcome is taken for both groups. If there is a difference in outcomes between the groups, it can reasonably be inferred that the difference was attributable to the policy.

Randomization creates the best comparisons because the two groups—treated and untreated—are more alike than in any other design. Because randomization, with sufficiently large numbers of units, creates

³For further discussion of these issues, see Gerring and McDermott (2007).

two groups in which all characteristics can be assumed to be equally distributed across the two groups, there is technically no need to have preintervention baseline measures, as these measures are assumed to be the same in each group due to their random assignment. The ability to do without baseline measures in large N randomized assignment designs could actually reduce the expenditure on this type of evaluation, as opposed to the costs incurred in other designs that require gathering data on baseline indicators. As discussed above, in the context of many projects in a country, gathering baseline data to evaluate the intervention in different ways, and measure other efforts, including activities and outputs would still be valuable.

Another advantage of randomized assignment in large N studies is that it often is perceived as the fairest method of distributing assistance in cases where the ability of USAID to provide DG assistance is limited and cannot cover all available units. Thus, for example, if only a certain fraction of judges or legislators in a country, or a certain fraction of villages in a district, can be served by a given assistance program, having a lottery to determine who gets assistance first is often judged even by participants as the fairest way to allocate resources. Since this method also creates the best impact evaluation design, it is a situation in which the ethics of assigning assistance and the goals of evaluation design are mutually reinforcing.

Common variations on the randomized treatment include “rollout” and “waiting list” protocols. With rollout protocols the treatment is given sequentially to different groups, with the order in which groups receive the treatment determined by random assignment. This solves the problem of how to distribute valued resources in a way that eventually makes them available to all but without destroying the potential for randomized control. It also offers the possibility of varying the treatment across each cohort, contingent on findings from previous cohorts. With waiting list protocols, the control group is comprised of those groups that are otherwise qualified and hence similar to the groups receiving treatment but were placed on a waiting list because of limits on funding. Evaluation is then undertaken on random samples from both the treatment and waiting list (control) groups. These latter groups may (or may not) be treated in subsequent iterations.

There are a number of well-known problems that can undermine the effectiveness of this research design, which can be found in many methodology texts (e.g., see Box 9.2 in Trochim and Donnelly 2007), some of which will be discussed here. Perhaps most noteworthy in the case of many USAID projects is the risk of contamination, in which the treatment of some individuals or groups (e.g., training some judges or legislators) also affects the behavior of those not enrolled in training. In addition,

randomized designs may encounter other problems, such as units refusing to participate in the design or units dropping out in the middle of the intervention. However, if large numbers of cases are available, most of these issues can be reasonably dealt with by amending the research design, so that if recognized and managed, these problems will not fatally undermine the validity of the evaluation.

The committee recognizes that political pressures to work with certain groups or locations, or to “just get the project rolling,” can work against the careful design and implementation of randomized assignments. These and other problems are addressed in a more detailed discussion of how to apply randomization to actual USAID DG projects in Chapter 6. The present chapter focuses mainly on the methodological reasons why the efforts needed to carry out randomized assignments for project evaluations can be worthwhile in terms of the increased confidence they provide that genuine causal relationships are being discovered and hence real project impact.

Unfortunately, from the standpoint of making the most credible impact evaluations, the units chosen to receive interventions from USAID are seldom selected at random. For example, nongovernmental organizations (NGOs) chosen for funding are often selected through a competition that results in atypical NGOs getting treatments. Or judges and administrators chosen to attend training workshops are selected based mainly on their willingness to participate. The problem here is that the criteria used for selecting NGOs and judges/administrators for participation in the project are almost certainly associated with a higher propensity to succeed in the project than would be the case for the “typical” NGO or judge, and this makes it impossible to assess project efficacy. If funded NGOs are found to do well or judges/administrators who attended workshops perform better, there is no way to rule out the possibility that the success observed is simply a function of having chosen groups or people who would have succeeded anyway or whose success was much greater than could generally be expected. The only way to avoid this pitfall—and to be in a position to know whether or not the project has had a positive impact—is to choose project participants randomly and then compare their performance to participants who were not selected to take part in the activity in question.

The bottom line is that if there is a strong commitment to answering the question—“Did the resources spent on a given project truly yield positive results?”—the best way to reach the most definitive answer is through an impact evaluation that involves the random selection of units for treatment and the collection of data in both treatment and control groups. As discussed in Chapter 6, many USAID DG projects that the committee encountered in the field were quite amenable in principle to randomiza-

tion without significant changes in their design. And it bears repeating that in some cases of large *N* randomized treatment, USAID may be able to eliminate the costs of collecting baseline data, which might make this evaluation design more attractive.

Randomized evaluations are useful for determining not only whether or not a given project/activity has had an effect but also where it appears to be most effective. To see this, consider Figure 5-1, which displays hypothetical data collected on outcomes among treatment and control groups for a particular USAID project. In this example, higher scores represent more successful outcomes.

Based on these data, it can be concluded that the treatment was a success since, on average, units (people, municipalities, courts, NGOs, etc.) given the treatment scored better on the outcome of interest than units in the control group. (This would need to be confirmed with a statistical test, but for now assume the two distributions are indeed different.) It is important to point out, however, that not every unit in the treatment group did better than units in the control group. Some units in the control group did better than those in the treatment group, and some in the treatment group did worse than those in the control group. In fact, at least a handful of units in the treatment group did worse than the average unit in the control group. Also, there is quite a bit of variance in the performance of those in the treatment group. By exploring the factors associated with high and low scores among the treatment cohort, inferences can be made about which ones predispose recipients of the treatment to success or fail-

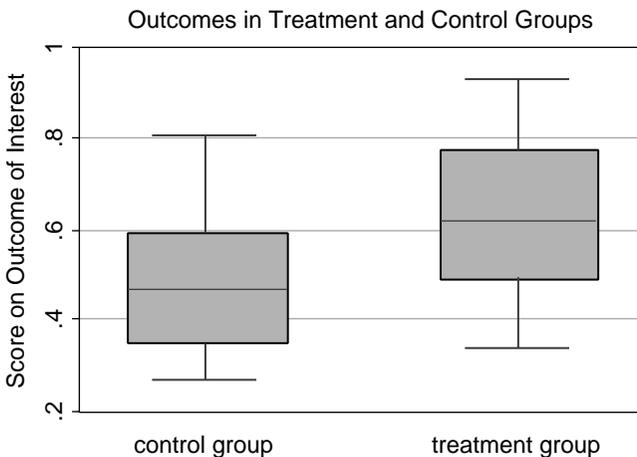


FIGURE 5-1 Hypothetical outcome data from treatment and control groups.

ure (or, put slightly differently, where the project works well and not so well). Thus the randomized design allows us to conclude not just whether the project was effective in achieving its goals but also where efforts should be directed in the next phase in order to maximize the impact.

2. Large N Comparison

Despite the utility of the large N randomized design, sometimes it is simply not possible to assign units randomly to the treatment group, even when the total number of units is large. The benefits of a large number of units for observing multiple iterations of the treatment, however, can still be exploited *if* one can overcome the following challenge: identifying and measuring those pretreatment differences between the treatment and control groups that might account for whatever posttreatment differences are observed. In these circumstances there are a variety of statistical procedures (e.g., propensity score matching, instrumental variables) for correcting the potential selection bias that complicates the analysis of causal effects.

The “matching” research design seeks to identify units that are similar to the ones getting treatment and then comparing outcomes.⁴ For example, Heckman et al (1997) sought to evaluate a jobs training project in the United States—the Job Training Partnership Act (JTPA). The JTPA provides on-the-job training, job search assistance, and classroom training to youth and adults who qualify (see Devine and Heckman [1996] for a more detailed analysis of the program). The U.S. Department of Labor commissioned an evaluation of the project to assess the impact of the main U.S. government training project for disadvantaged workers. Evaluators collected longitudinal data on those individuals who went through the JTPA and those who did not. Since the individuals who received the services were not chosen randomly, the evaluators constructed a nontreated group to compare them with, based on a number of criteria that matched the “in group” along many characteristics, such as location of residence, eligibility for the program, income, and education. Using this matching design, the evaluators were able to compare the effect of the project by gathering data before and after it started.

Another technique to use in a large N situation is the regression discontinuity design (Shadish et al 2002:Chap. 7, Hahn et al 2001). Regression discontinuity is used in situations where the assignment of the treatment is based on the characteristics of the group that a policy is designed to affect, and the before and after outcomes of interest are measured for both groups. For example, in a reading program the assignment of a remedial

⁴See Heckman (1997) for a more extensive discussion of the implicit behavioral assumptions that justify the method of matching.

reading project is based on the preproject tests on the readers. At some cutoff point, students are assigned to the project or not. The expectation is that project success would produce a more positive trend after the intervention for those below the cutoff point. The trend before and after the intervention is estimated, and the differences are compared to see if the intervention had any discernible effect.

Angrist and Lavy (1999), for example, used the regression discontinuity design to evaluate the effect of classroom size on student test scores in Israel. They compared classes with greater than and less than 40 students and found that class size was, in fact, linked to test performance.

Yet another design useful to large N samples is the difference-in-difference (DD) approach. A DD design compares two cases, one that received the project and one that did not and compares the difference between their before and after levels on the relevant outcome variable. DD estimation has become a widespread method to estimate causal relationships (Bertrand et al 2004). For example, if the DG project provides assistance to one judge and not another, before and after measures of a particular outcome variable should be taken for both and compared. In a regression that followed this design, the differences for each judge's behavior and between each judge's behavior are both estimated. The appeal of DD comes from its simplicity as well as its potential to circumvent many of the endogeneity problems that typically arise when making comparisons between heterogeneous individuals (Meyer et al 1995).

In an example of this approach, Duflo (2000) used a DD design to evaluate the effect of school construction on education and wages in Indonesia. Across several regions she compares one region's school construction with another that has not yet had its construction. As always, baseline data were critical to discovering any effect from the program. This design is useful when there is only one or a few treated units and is better than just a before-and-after analysis of a single unit since it offers a controlled comparison.

Efforts to use statistical methods to approximate randomized designs are only as effective as the evaluator's ability to model the selection process that led some units to be given the treatment while others were not. Attention to gathering a battery of pretreatment measures across cases is critical to an effective large N comparison. With sufficient cases and systematic efforts to measure pre- and posttreatment outcomes, large N comparisons can provide meaningful insights into project impacts even when the treatment cannot be manipulated through randomization by USAID.

3. Small N Randomization

In some instances it is possible to manipulate the policy of interest

(the treatment) but only across a very small set of cases. In this case it is not possible to use probability tests derived from statistical theory to gauge the causal impact of an experiment across groups where the treatment and control groups each have only one or several members or where there is no control whatsoever. However, in other respects the challenges posed by, and advantages accrued from, this sort of analysis are quite similar to the large N randomized design.

Where cross-unit variance is minimal (by reason of the limited number of units at one's disposal), the emphasis of the analysis necessarily shifts from spatial evidence (across units) to evidence garnered from temporal variation (i.e., to a comparison of pre- and posttests in the treated units). Naturally, one wants to maximize the number of treated units and the number of untreated controls. This can be achieved by a modified "rollout" protocol. Note that in a large N randomized setting (as described above), the purpose of rollout procedures is usually (1) to test a complex treatment (e.g., where multiple treatments or combinations of treatments are being tested in a single research design) or (2) for purposes of distributing a valued good among the population while preserving a control group. The most crucial issue is to maximize useful variation on the available units. This can be achieved by testing each unit in a serial fashion, regarding the remaining (untreated) units as controls.

Consider a treatment that is to be administered across six regions of a country. There are only six regions, so cross-unit variation is extremely limited. To make the most of this evidence-constrained setting, the researcher may choose to implement five iterations of the same manipulated treatment, separated by some period of time (e.g., one year). During all stages of analysis, there remains at least one unit that can be regarded as a control. This style of rollout provides five pre- and posttests and a continual (albeit shrinking) set of controls. As long as contamination effects are not severe, the results from this sort of design may be more easily interpreted than the results from a simple split-sample research design (i.e., treating three regions and retaining the others as a control group). In the latter any observed variation across treatment and control groups may be due to a confounding factor that coincides temporally and correlates spatially with the intervention.

Despite the randomized nature of this intervention, it is still quite possible that other matters beyond the control of the investigator may intercede. It is not always possible to tell whether or not confounding factors are present in one or more of the cases. In a large N setting, we can be more confident that such confounding factors, if present, will be equally distributed across treatment and control groups. Not so for the small N setting. This is all the more reason to try to maximize experimental leverage by setting in motion a rollout procedure that treats each unit

separately through time. Any treatment effects that are fairly consistent across the six cases are unlikely to be the result of confounding factors and are therefore interpretable as causal rather than spurious.

Note that in a small population where all units are being treated, it is likely that there will be significant problems of contamination across units. In the scenario discussed above, for example, it is likely that untreated regions in a country will be aware of interventions implemented in other regions. Thus it is advisable to devise case selection and implementation procedures that minimize potential contamination effects. For example, in the rollout protocol discussed above, one might begin by treating regions that are most isolated, leaving the capital region for last.

Regardless of the procedure for case selection, it will be important for researchers to pay close attention to potential changes before and after the treatment is administered. That is, in small N randomization designs, it is highly advisable to collect baseline data since the comparison groups are less likely to be similar enough to compare directly.

In an example of a small N randomized evaluation, Glewwe et al (2007) used a very modest sample of 25 randomly chosen schools to evaluate the effect of the provision of textbooks on student test scores. A Dutch nonprofit organization provided textbooks to 25 rural Kenyan primary schools chosen randomly from a group of 100 candidate schools. The authors found no evidence that the project increased average scores, reduced grade repetition, or affected dropout rates (although they did find that the project increased the scores of the top two quintiles of those with the highest preintervention academic achievement). Evidently, simply providing the textbooks only helped those who were already the most motivated or accomplished; in the absence of other changes (e.g., better attendance, more prepared or involved teachers), the books alone produced little or no change in average students' achievement. It is important to note that, like other forms of impact evaluation, this study required good baseline data to conduct its evaluation.

4. Small N Comparison

In small N designs USAID may be unable to manipulate the temporal or spatial distribution of the treatment. In this context the evaluator faces the additional hurdle of not having sufficient cases to employ statistical procedures to correct for the biases that make identifying causal effects difficult when treatments cannot be manipulated.

Nonetheless, there are still advantages to identifying units that will *not* be treated and gathering pre- and posttreatment measures of outcomes in both the treatment and control groups. A control group is useful here for (1) ruling out the possibility that the intervention coincided with a temporal change or trend that might account for observed changes in

the treatment group and (2) ensuring that application of the treatment was not correlated with other characteristics of the treated units that could explain observed differences between the treatment and control groups. Ideally, the control group in a small N comparison should be matched to the treatment group as precisely as possible. With large amounts of data, propensity score matching techniques can be used to identify a control group that approximates the treated units across a range of observables. When data are not widely available, a control group can be generated qualitatively by identifying untreated units that are similar to those in the treatment group on key dimensions (other than the treatment) that might affect the outcomes of interest.

5. $N = 1$ Study with USAID Control over Timing and Location of Treatment

Sometimes, there is no possibility of spatial comparison. This is often the case where the unit of concern exists only at a national level (e.g., an electoral administration body), and nearby nation-states do not offer the promise of pre- or posttreatment equivalence. In this case the researcher is forced to reach causal inferences on the basis of a single case. Even so, the possibility of a manipulated treatment offers distinct advantages over the unmanipulated (observed) treatment. The ability to choose the timing of the intervention and plan observations to maximize the likelihood of accurate inferences can provide considerable leverage for credible conclusions. However, these advantages accrue only if very careful attention is paid to the timing of the intervention, the nature of the intervention, its anticipated causal effect, and the pre- and posttreatment evidence that might be available. The challenge here is to overcome the problems that are already highlighted here with regard to simple before and after comparisons.

First, with respect to timing, it is essential that the intervention occur during a period in which no other potentially contaminating factors are at work and in which the outcome factors being observed would be expected to be relatively stable; that is, a constant trend is expected, so that any changes in that trend are easily interpreted. Naturally, these matters lie partly in the future and therefore cannot always be anticipated. Nonetheless, the delicacy of this research design—its extreme sensitivity to any violation of *ceteris paribus* assumptions—requires the researcher to anticipate what may occur, at least through the duration of the experiment.

Second, with respect to detrending the data, it is helpful if the researcher can gather information on the outcome(s) of interest and any potential confounders for the periods before and after the intervention. The longer the period of observation, the more confident one can be about any causal inference made (Campbell 1968/1988). Thus, if the outcome

factor being studied has been stable for a long time before the intervention, and other factors likely to have an impact on the outcome have been ruled out, one can have more confidence that any observed change in the trend was due to the intervention.

Third, with respect to the intervention itself, it is essential that it be discrete and significant enough to be easily observed. While subtle project effects may be detected in a large N randomized design, usually only very large effects can be confidently observed in a single-case setting.

Fourth, it is helpful if the intervention has more than one observable (and policy-significant) effect. This goes some way toward resolving the ever-present threats of measurement error and confounding causes. If, for example, a given intervention is expected to produce changes in three measurable independent outcomes, and all three factors change in the aftermath of an intervention, it is less likely that the noted association is spurious.

6. $N = 1$ Comparison

When the unit of concern exists only at the national level and the treatment cannot be manipulated by USAID, discerning causal effects is extraordinarily difficult. Observed differences in outcome measures pre- and posttreatment can be interpreted as causal effects only if the evaluator can make the case that other factors were not important.

Some of the strategies described above are applicable in an $N = 1$ comparison if the treatment can be interpreted "as if" it was manipulated (e.g., Miron 1994). Any demonstration of a large discontinuous change in an outcome of interest following the treatment increases confidence in the causal interpretation of the effect. This requires an effort to measure the outcome(s) of interest prior to, and after, the intervention.

In some cases it may be possible to identify units for comparison *within* the country or *outside* the country, in order to rule out obvious temporal confounds. Take the example of an anticorruption effort funded in a specific ministry. If it can be shown that corruption levels remained unchanged in untreated ministries while shifting dramatically in a treated ministry, we gain confidence that a government-wide anticorruption effort cannot account for the effects observed in the treated ministry. But the possibility cannot be ruled out that other developments in the treated ministry (such as good leadership) are more important than the intervention in accounting for the outcome. Or take the example of a national anticorruption effort that is rolled out in one country but not in adjacent countries or at different times in adjacent countries. Changes in outcome variables in the other countries could be tracked to seek the effects of the program; if reductions in corruption occur to a greater degree, or in a timed sequence that corresponds to the timing of roll-outs in different

countries, one can have confidence that it is not regional or global trends that were driving the reductions in corruption. On the other hand, as in the previous example, the possibility could not be ruled out that other factors, such as freer media or stronger leadership, were the key causal factors in reducing corruption rather than the specific USAID project, unless there were also measures of those possible confounding factors.

Not all USAID DG programs need to be subjected to rigorous impact evaluation. For example, if USAID is working to help a country pass a new constitution with certain human rights provisions, and several other NGOs and foreign countries are also working to that end, it may not matter how much USAID's specific activities contributed to a successful outcome; success is what matters and credit can be shared among all who contributed. (On the other hand, a subsequent impact evaluation of whether the new constitution actually resulted in an improvement in human rights—an $N = 1$ comparison designed to plot changes in human rights violations over time and look for sharp reductions following adoption of the new constitution—may be worthwhile.)

In particular, the random assignment mode of impact evaluation is probably best used only where the fair assignment of assistance naturally results in a randomized assignment of aid or where USAID uses a project in so many places, or invests so much in a project, that it is of great importance to be confident of that project's effectiveness. In most settings, worthwhile insights into project impacts can be derived from designs that include small N comparisons, as long as good baseline, outcome, and comparison group data are collected.

EXAMPLES OF THE USE OF RANDOMIZED EVALUATIONS IN IMPACT EVALUATIONS OF DEVELOPMENT ASSISTANCE (INCLUDING DG PROJECTS)

Randomized designs have a high degree of internal validity. By permitting a comparison of outcomes in a treatment group and a control group that can be considered identical to one another, they do a better job than any other evaluation technique of permitting evaluators to identify the impact of a given intervention. It is no surprise, therefore, that randomized evaluation is the industry standard for the assessment of new medications. It is inconceivable that a pharmaceutical company would be permitted to introduce a new medication into the market unless evidence from a randomized evaluation proved its benefits. Yet as discussed in Chapter 2, for the assessment of DG assistance programs, impact evaluations have rarely been employed. This leaves USAID in the difficult position of spending hundreds of millions of dollars on assistance programs without proven effects.

There are a small, but important, number of large N randomized impact evaluations that have been carried out to test the effects of assistance programs. Classic evaluations, such as the RAND health insurance study and the evaluation of the Job Training Partnership Act (JTPA), stand out as exemplars of large-scale assessments of social assistance programs (Wilson 1998, Gueron and Hamilton 2002, Newhouse 2004). A few have been done in developing countries; the evaluation of Mexico's conditional cash transfer program, Progresa/Oportunidades, continues to shape the design of similar programs in other contexts (Morley and Coady 2003).

The number of such evaluations is growing. In fields as diverse as public health, education, microfinance, and agricultural development, randomized evaluations are increasingly employed to assess project effectiveness. Examples abound in the field of public health: Studies have assessed the efficacy of male circumcision in combating HIV (Auvert et al 2005), the impact of HIV prevention programs on sexual behavior (Dupas 2007), the effectiveness of bed nets for reducing the incidence of malaria (Nevill et al 1996), the impact of deworming drugs on health and educational achievement (Miguel and Kremer 2004), and the role of investments in clean water technologies on health outcomes (Kremer et al 2006). In education, randomized evaluations have been used to explore the efficacy of conditional cash transfers (Schultz 2004), school meals (Vermeersch and Kremer 2004), and school uniforms and textbooks (Kremer 2003) on school enrollment; the effectiveness of additional inputs, such as teacher aids, on school performance (Banerjee and Kremer 2002); and the impact of school reforms, such as voucher programs, on academic achievement (Angrist et al 2006). In microfinance, attention has focused on the impact of programs on household welfare (Murdoch 2005); randomized evaluations in agricultural development are exploring the benefits and impediments to the adoption of new technologies, such as hybrid seeds and fertilizer (Duflo et al 2006a).

Thus far, however, these approaches have not been applied to the evaluation of DG programs. A significant part of the explanation for this is that it is often more difficult to measure outcomes in the area of democratic governance. Most successful randomized evaluations have been conducted in areas such as health and education, where it is much more straightforward to measure outcomes. For example, the presence of intestinal parasites can be measured quite easily and accurately via stool samples (as in Miguel and Kremer 2004); water quality can be assessed via a test for *E. coli* content (as in Kremer et al 2006); nutritional improvements can be traced quite readily via height and weight measures; school performance or learning can be tracked easily via test scores (as in Banerjee et al 2007); and teacher absenteeism can be measured with attendance records (as in Banerjee and Duflo 2006). Developing valid and reliable measures

of the outcomes targeted by DG programs is much more difficult and stands as an important challenge for project evaluation in this area. The challenge is not insurmountable; there have been tremendous improvements over the past decade in the measurement of political participation and attitudes (Bratton et al 2005), social capital and trust (Grootaert et al 2004), and corruption (Bertrand et al 2007, Olken 2007). And as discussed in Chapter 2, USAID has made significant efforts to develop outcome indicators to support its project M&E work.

This chapter closes with two examples of impact evaluations using randomized designs applied to DG subjects that tested commonly held programming assumptions. The first addresses the issue of corruption. USAID invests significant resources every year in anticorruption initiatives, but questions remain about the efficacy of such investments. Which programs yield the biggest impact in terms of reducing corruption? Some have argued that corruption can be reduced with the right combination of monitoring and incentives provided from above (Becker and Stigler 1974). Of course, the challenge with top-down monitoring is that higher level officials may themselves be corruptible. An alternative approach has emphasized local-level monitoring (World Bank 2004). The argument is that community members have the strongest incentives to police the behavior of local officials, as they stand to benefit the most from local public goods provision. Yet this strategy also has its drawbacks: Individuals may not want to bear the costs of providing oversight, preferring to leave that to others, or community members may be easily bought off by those engaged in corrupt practices. Which strategy most effectively reduces corruption?

Olken (2007) set out to answer this question in Indonesia through a unique partnership with the World Bank. As a nationwide village-level road-building project was rolled out, Olken randomly selected one set of villages to be subject to an external audit by the central government, a second set in which extensive efforts were made to mobilize villagers to participate in oversight and accountability meetings, a third set in which the accountability meetings were complemented by an anonymous mechanism for raising complaints about corruption in the project, and a fourth set reserved as a control group. To measure the efficacy of these different strategies, Olken constructed a direct measure of corruption: He assembled a team of engineers and surveyors who, after the projects were completed, dug core samples in each road to estimate the quantity of materials used, interviewed villagers to determine the wages paid, and surveyed suppliers to estimate local prices to construct an independent estimate of the cost of the project. The difference between the reported expenditures by the village and this independent estimate provides a direct measure of corruption. His findings strongly suggest the efficacy of

external audits: Missing expenditures were eight percentage points lower in villages subject to external monitoring. The results were less impressive for grassroots monitoring. While community members did participate in the accountability meetings in higher numbers in villages where special mobilization efforts were undertaken and they did discuss corruption-related problems (and even took action at times), no significant reductions in the level of corruption were observed. If one had relied on only the output measures or observation that characterizes many USAID M&E efforts (e.g., number of participants in community events supported by USAID programs), it might have mistakenly been concluded from the level of community participation that grassroots monitoring was making a substantial difference. But Olken's more careful methodology led him to the opposite conclusion. While there are undoubtedly benefits to mobilizing community participation for a variety of other purposes, it appears that if the goal is to reduce local corruption, supporting more external audits is considerably more effective.

Another example is the question of how best to promote a robust and vibrant civil society. USAID regularly makes substantial investments in civil society organizations (CSOs) and local NGOs with the hope of empowering the disadvantaged, building trust, enhancing cooperation, and supporting the flourishing of democratic institutions (Putnam 1993, 2000). Yet some skeptics have warned that outside support for CSOs might be counterproductive: It may produce more professionally run organizations that no longer have strong ties to their grassroots base (Skocpol 2003) and may actually change the leadership of such organizations, disempowering the disadvantaged (Igoe 2003). Knowing whether outside assistance helps or harms CSOs is a question of vital importance, and randomized evaluations have begun to offer some preliminary evidence.

Gugerty and Kremer (2006) conducted a randomized evaluation in which a sample of women's self-help associations in rural Western Kenya were randomly selected to receive a package of assistance that included organizational and management training as well as a set of valuable agricultural inputs such as tools, seeds, and fertilizer. Forty groups received assistance in the first year, while an additional 40 eligible groups served as the control group (although they were given the same assistance, just two years later). The results are disturbing for advocates of outside funding to community groups. While members of the funded groups reported higher levels of satisfaction with their group leadership, there is little evidence that objective measures of group activity improved. Moreover, Gugerty and Kremer found that outside funding changed the nature of the group and its leadership. Younger, more educated women and women from the formal sector increasingly joined the group, and these new entrants tended to assume leadership positions and to displace older women.

Compared to their unfunded counterparts, funded groups experienced a two-thirds increase in the exit rate of older women—a troubling finding given the program’s underlying objective of empowering the disempowered. Whereas an analysis of group members’ satisfaction would have led project evaluators to conclude that the project was a success, the careful randomized design led Gugerty and Kremer to the opposite conclusion (and generated significant evidence that the skeptics may be right about the sometimes counterproductive impact of donor funding to CSOs).

These two examples serve to support a broader point: It is both possible and important to conduct randomized impact evaluations of projects designed to support DG. In both cases the randomized evaluations effectively measured a project’s impact, but they also provided new evidence about implicit hypotheses that guide programming more broadly. In the case of corruption, the implicit hypothesis was that community empowerment is an antidote to local-level corruption; in the case of civil society support, the hypothesis was that donors can spur the growth of a vibrant civil society that empowers the disadvantaged through outside support. The evidence casts some doubt on both hypotheses and should encourage further evaluations to see if these results hold more broadly and perhaps fuel the search for alternative methods to support DG goals.

The larger point, however, is not so much the findings of these studies as the fact that they were successfully conducted on DG projects. The next chapter describes the findings of the committee’s field studies and discusses how these designs could be applied to the evaluation of several of USAID’s own current DG projects. It also explicitly addresses some of the common objections to using randomized evaluations more widely.

REFERENCES

- Angrist, J.D., and Lavy, V. 1999. Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics* 114:533-575.
- Angrist, J.D., Bettinger, E., and Kremer, M. 2006. Long-Term Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia. *American Economic Review* 96:847-862.
- Auvert, B., Taljaard, D., Lagarde, E., Sobngwi-Tambekou, J., Sitta, R., and Puren, A. 2005. Randomized, Controlled Intervention Trial of Male Circumcision for Reduction of HIV Infection Risk: The ANRS 1265 Trial. *PLOS Medicine*. Available at: <http://medicine.plosjournals.org/periserv/?request=get-document&doi=10.1371/journal.pmed.0020298&ct=1>. Accessed on February 23, 2008.
- Banerjee, A.V., and Duflo, E. 2006. Addressing Absence. *Journal of Economic Perspectives* 20(1):17-132.
- Banerjee, A.V., and Kremer, M. 2002. Teacher-Student Ratios and School Performance in Udaipur, India: A Prospective Evaluation. Washington, DC: Brookings Institution.
- Banerjee, A.V., Cole, S., Dutlo, E., and Linden, L. 2007. Remediating Education: Evidence from Two Randomized Experiments in India. *Quarterly Journal of Economics* 122(3):1235-1264.

- Becker, G.S., and Stigler, G.J. 1974. Law Enforcement, Malfeasance, and the Compensation of Enforcers. *Journal of Legal Studies* 3:1-19.
- Bertrand, M., Duflo, E., and Mullainathan, S. 2004. How Much Should We Trust Difference-In-Difference Estimates? *Quarterly Journal of Economics* 119(1):249-275.
- Bertrand, M., Djankov, S., Hanna, R., and Mullainathan, S. 2007. Obtaining a Driving License in India: An Experimental Approach to Studying Corruption. *Quarterly Journal of Economics* 122:1639-1676.
- Bloom, H.S., ed. 2005. *Learning More from Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage Foundation.
- Bollen, K., Paxton, P., and Morishima, R. 2005. Assessing International Evaluations: An Example from USAID's Democracy and Governance Programs. *American Journal of Evaluation* 26:189-203.
- Bratton, M., Mattes, R., and Gyimah-Boadi, E. 2005. *Public Opinion, Democracy, and Market Reform in Africa*. New York: Cambridge University Press.
- Campbell, D.T. 1968/1988. The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis. Pp. 222-238 in *Methodology and Epistemology for Social Science*, E.S. Overman, ed. Chicago: University of Chicago Press.
- Devine, T.J., and Heckman, J.J. 1996. The Economics of Eligibility Rules for a Social Program: A Study of the Job Training Partnership Act (JTPA)—A Summary Report. *Canadian Journal of Economics* 29(Special Issue: Part 1):S99-S104.
- Duflo, E. 2000. Schooling and Labor Market Consequences for School Construction in Indonesia. Cambridge, MA: MIT Dept. of Economics Working Paper No. 00-06.
- Duflo, E., Kremer, M., and Robinson, J. 2006a. Understanding Technology Adoption: Fertilizer in Western Kenya. Evidence from Field Experiments. Available at: http://www.econ.berkeley.edu/users/webfac/saez/e231_s06/esther.pdf. Accessed February 23, 2008.
- Duflo, E., Glennerster, R., and Kremer, M. 2006b. Using Randomization in Development Economics Research: A Toolkit. Unpublished paper. Massachusetts Institute of Technology and Abdul Latif Jameel Poverty Action Lab. Cambridge, MA.
- Dupas, P. 2007. Relative Risks and the Market for Sex: Teenagers, Sugar Daddies, and HIV in Kenya. Available at: <http://www.dartmouth.edu/~pascaline/>. Accessed on February 23, 2008.
- Gerring, J. 2007. *Case Study Research: Principles and Practices*. Cambridge: Cambridge University Press.
- Gerring, J., and McDermott, R. 2007. An Experimental Template for Case-Study Research. *American Journal of Political Science* 51:688-701.
- Glewwe, P., Kremer, M., and Moulin, S. 2007. Many Children Left Behind? Textbooks and Test Scores in Kenya. NBER Working Paper No. 13300. Available at: <http://papers/nber.org/papers/w13300>. Accessed on April 26, 2008.
- Gueron, J.M., and Hamilton, G. 2002. The Role of Education and Training in Welfare Reform. Policy Brief No. 20. Washington, DC: Brookings Institution.
- Grootaert, C., Narayan, D., Jones, V.N., and Woolcock, M. 2004. Measuring Social Capital: An Integrated Questionnaire. World Bank Working Paper No. 18. Washington, DC: The World Bank.
- Gugerty, M.K., and Kremer, M. 2006. Outside Funding and the Dynamics of Participation in Community Associations. Background Paper. Washington, DC: World Bank. Available at <http://siteresources.worldbank.org/INTPA/Resources/Training-Materials/OutsideFunding.pdf>. Accessed on April 26, 2008.
- Hahn, J., Todd, P., and Van der Klaauw, W. 2001. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica* 69(1):201-209.
- Heckman, J. 1997. Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations. *Journal of Human Resources* 32(3):441-462.

- Heckman, J., Ichimura, H., and Todd, P. 1997. Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program. *Review of Economic Studies* 64:605-654.
- Hyde, S. 2006. The Observer Effect in International Politics: Evidence from a Natural Experiment. Unpublished paper. Yale University.
- Igoe, J. 2003. Scaling Up Civil Society: Donor Money, NGOs and the Pastoralist Land Rights Movement in Tanzania. *Development and Change* 34(5):863-885.
- Kremer, M. 2003. Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons. *American Economic Review* 93(2):102-106.
- Kremer, M., Leino, J., Miguel, E., and Zwane, A. 2006. Spring Cleaning: A Randomized Evaluation of Source Water Improvement. Available at <http://economics.harvard.edu/faculty/Kremer/files/springclean.pdf>. Accessed on April 26, 2008.
- Meyer, B.D., Viscusi, W.K., and Durbin, D.L. 1995. Workers' Compensation and Injury Duration: Evidence from a Natural Experiment. *American Economic Review* 85(3):322-340.
- Miguel, E., and Kremer, M. 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* 72(1):159-217.
- Miron, J.A. 1994. Empirical Methodology in Macroeconomics: Explaining the Success of Friedman and Schwartz's "A Monetary History of the United States, 1867-1960." *Journal of Monetary Economics* 34:17-25.
- Morley, S., and Coady, D. 2003. *Targeted Education Subsidies in Developing Countries: A Review of Recent Experiences*. Washington, DC: Center for Global Development.
- Murdoch, J. 2005. *The Economics of Microfinance*. Cambridge, MA: MIT Press.
- Nevill, C.G., Some, E.S., Mung'ala, V.O., Mutemi, W., New, L., Marsh, K., Lengeler, C., and Snow, R.W. 1996. Insecticide-Treated Bednets Reduce Mortality and Severe Morbidity from Malaria Among Children on the Kenyan Coast. *Tropical Medicine and International Health* 1:139-146.
- Newhouse, J.P. 2004. Consumer-Directed Health Plans and the RAND Health Insurance Experiment. *Health Affairs* 23(6):107-113.
- Olken, B.A. 2007. Monitoring Corruption: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy* 115:200-249.
- Putnam, R. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton, NJ: Princeton University Press.
- Putnam, R. 2000. *Bowling Alone*. New York: Simon and Schuster.
- Schultz, T.P. 2004. School Subsidies for the Poor: Evaluating the Mexican PROGRESA Poverty Program. *Journal of Development Economics* 74(1):199-250.
- Senge, P. 2006. *The Fifth Discipline: The Art and Practice of the Learning Organization*, Rev. Ed. New York: Doubleday.
- Skocpol, T. 2003. Diminished Democracy: From Membership to Management in American Civic Life. Julian T. Rothbaum Lecture Series, Vol. 8. Norman: University of Oklahoma Press.
- Shadish, W.R., Cook, T.D., and Campbell, D.T. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.
- Trochim, W., and Donnelly, J.P. 2007. *The Research Methods Knowledge Base*, 3rd ed. Cincinnati, OH: Atomic Dog Publishing.
- Vermeersch, C., and Kremer, M. 2004. School Meals, Educational Achievement, and School Competition: Evidence from a Randomized Evaluation. World Bank Policy Research Working Paper No. 3523. Washington, DC: The World Bank.
- Wholey, J.S., Hatry, H.P., and Newcomer, K.E., eds. 2004. *Handbook of Practical Program Evaluation*, 2nd ed. San Francisco: Jossey-Bass.
- Wilson, E.O. 1998. *Consilience: The Unity of Knowledge*. New York: Knopf.
- World Bank. 2004. *Monitoring & Evaluation: Some Tools, Methods & Approaches*. Washington, DC: International Bank for Reconstruction, The World Bank.

6

Implementing Impact Evaluations in the Field

INTRODUCTION

A counterfactual question—how would things have looked in the absence of the U.S. Agency for International Development (USAID) program?—lies at the core of any design for impact evaluations. Chapter 5 made the case that randomized evaluations provide the soundest methodology for generating definitive answers to this question. However, it is one thing to specify what may be optimal theoretically and another thing altogether to implement that methodology on the ground. Practical impediments may make the implementation of randomized evaluation difficult, even impossible, at least in a pure form. For example, factors outside of USAID’s control may render it not feasible to gather baseline data, to identify and monitor outcomes in a control group, or to select by lottery the units in which programs should be implemented. Although Chapter 5 provided examples of several successful randomized evaluations, only a handful of these are in the democracy and governance (DG) area, and none of them are examples of evaluations of USAID’s own programs. Thus, even if willing to accept the desirability in principle of adopting the methodology of randomized evaluation, it is reasonable to wonder how readily it can be applied to the sorts of programs that USAID missions in the field regularly undertake.

To find out, the committee commissioned three expert teams to visit USAID missions overseas in an effort to assess the viability of impact evaluations for past and present DG programming. The key task for each team was to talk with implementers, local partners, and USAID mission

personnel on the ground to assess the feasibility of actually implementing in practice the evaluation methodologies outlined in the previous chapter. The first part of this chapter presents the results of those field visits. The second part provides responses to the most commonly raised objections that the committee and its field teams heard expressed about the use of randomized evaluations in DG programs.

Before turning to the details of what the field teams found, it is important to highlight a clear and consistent message that came through from all three field visits. All three teams concluded, first, that *the introduction of randomized evaluations into USAID project evaluation was both feasible and cost-effective in many of the contexts they investigated*. They were unanimous that, where possible, adopting such methods would represent an improvement over current practices. Second, they reported that, *for projects where randomized evaluations were not possible, other improvements to USAID evaluation—for example, improved measurement, systematic collection of baseline data, and comparisons across treated and untreated units—also have the potential to yield significant improvements in the agency’s ability to attribute project impact*. These issues are discussed in Chapter 7. Finally, the teams returned from the field energized by their interactions with mission staff and confident that a willingness, and even excitement, exists about improving the quality of project evaluations. The teams were also impressed with some of the work already being done as part of current project monitoring, in particular in the broadening of measurement strategies beyond project outputs to include an assessment of outcomes.

FIELD VISITS TO USAID MISSIONS

As a complement to the deliberations in Washington and extensive engagement with USAID staff and implementers, the committee felt strongly that its recommendations should be informed by a set of extended field visits to USAID missions. The committee therefore identified a set of missions, representing a diversity of regions, that were engaged in substantial programming on DG issues and were in the process of designing large, new projects in one of USAID’s core DG areas (rule of law, elections and political processes, civil society, and governance). From the list of missions provided, USAID explored the willingness of the missions to host the team and consider new approaches to project evaluation. After negotiating issues of timing and access, USAID and the committee agreed to send field teams to Albania, Peru, and Uganda. The field visits were intended to accomplish three main goals:

1. to better understand current strategies used for project evaluation, including approaches to data collection;

2. to explore the feasibility of introducing impact evaluations in the future, including (but not limited to) randomized evaluation; and
3. to obtain the perspectives of mission personnel and USAID implementers regarding the possibilities for, and impediments to, new approaches to evaluation.

The committee encouraged the field teams to explore the range of DG activities currently under way in each mission, assess the adequacy of current evaluation approaches, and provide concrete examples of how existing approaches could be improved. In addition, the field teams were directed to focus particular attention on the development of an impact evaluation design in one specific area in each mission. The teams focused on local government/decentralization in Albania and Peru and support for multiparty democracy in Uganda.¹

Each field team was composed of methodological consultants, academic or other experts with relevant experience in research design or program evaluation and DG issues, and country or regional expertise; a Washington-based USAID staff member who was familiar with the mission, the committee's work, and USAID policies and practices; and National Research Council professional staff, who assisted the consultants in meeting the team's objectives and coordinated the logistics of the field visits.

In evaluating the findings of the three field teams, it is important to keep in mind that the field teams visited missions that had expressed an interest in improving their evaluation strategies. The field teams' conclusions about the applicability of impact evaluations, especially its sense that standard objections to these designs can be addressed, thus reflect the experiences gleaned from this (nonrandom) sample of missions. It is not known if other missions, especially smaller ones with leaner budgets or those in countries experiencing violent conflicts or particularly rapid political change, would be as amenable to new approaches to evaluation: The committee has no control group of non-self-selecting missions with which to compare its findings. Yet the committee believes it unlikely that missions that did not invite the committee to send a field team would have offered novel additional objections. Over the 15 months of the study period, the committee talked with numerous USAID staff and implementers from a variety of areas and with backgrounds and experience with DG programming in a great many countries, and the set of objections that are taken up in the second part of this chapter dominated the responses of everyone with whom the committee spoke.

¹Key results of the field visits are discussed in this chapter and the next. Additional information can be found in Appendix E.

EMPLOYING RANDOMIZED IMPACT EVALUATIONS FOR USAID DG PROJECTS IN THE FIELD

Randomized evaluations are widely considered the best method for determining the causal effects of treatment in a broad range of areas, including public health, education, microfinance, and agriculture. As the Olken (2007) and Gugerty and Kremer (2006) studies described in Chapter 5 show, such methodologies are also beginning to be applied to evaluate the effectiveness of projects in the area of democratic governance. Nonetheless, the committee learned from its consultations with USAID staff and implementers that there is a general feeling that randomized evaluation was not an option for many of the projects that USAID carries out. Even in those cases where randomized evaluations might be possible theoretically, the assumption among USAID staff seemed to be that such approaches would be too difficult to implement in practice, owing to an inability to select treatment groups by lottery, the difficulty of preserving a control group, the difficulty of identifying good indicators for key outcomes, the high cost of the extensive data collection that would be required, or the tension between the flexibility staff believe they need to respond to opportunities and challenges as projects go forward and the need to minimize changes to ensure an effective evaluation.

These are legitimate concerns. To address them, this section discusses how randomized evaluations could be used in current USAID projects, drawing on examples gleaned from the field visits and consultations with practitioners. We begin with a decentralization project in Peru that has already been implemented, outlining how the project monitoring strategy that was employed could have been adjusted to accommodate a randomized component that would have made it an impact evaluation design and showing how such an adjustment would have permitted the mission to generate much stronger inferences about project impact.² Then a planned multipronged effort to support multiparty democracy in Uganda is described, emphasizing how pieces of the existing project might be amenable to randomized evaluation and showing how adopting such an evaluation method would improve USAID's ability to assess the project's effects.³ The committee's goal is to use these projects as illustrations of the potential payoffs that could accrue from improved evaluation strategies.

²The discussion here of decentralization in Peru is drawn from the report of a field team led by Thad Dunning, assistant professor of political science, Yale University.

³These designs were developed by a team led by Devra Moehler, assistant professor of political science, Cornell University.

Decentralization in Peru

USAID/Peru launched a project 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 subnational governments (such as “participatory budgeting”),
- strengthen the management skills of subnational governments in selected regions of Peru, and
- increase the capacity of nongovernmental 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 project took place in seven selected subnational regions (also called departments): Ayacucho, Cusco, Huanuco, Junin, Pasco, San Martin, and Ucayali.⁴ These seven regions contain 61 provinces, which in turn contain 536 districts.⁵ Workshops on participatory budgeting, training of civil society organizations (CSOs), and other interventions took place at the regional, provincial, and district levels.⁶

The ultimate goal of the project was to promote “increased responsiveness of subnational elected governments to citizens at the local level in selected regions” (USAID/Peru 2002). This outcome is potentially measurable on different units of 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).

The PRODES decentralization project 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 in relation to project outputs or very proximate outcomes, which

⁴The regions were nonrandomly selected for programs because they share high poverty rates, significant indigenous populations, and narcotics-related activities and because a number of the departments were strongholds for the Shining Path movement in the 1980s.

⁵Peru 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.

⁶Relevant subnational authorities include members of regional councils, provincial mayors, and mayors of districts.

were the focus of the project monitoring plan used by the implementer.⁷ Instead, the question is how we could know whether such a project had impacts on targeted policy outcomes, such as the responsiveness of local governments to citizens' demands.

Since the project was not designed with impact evaluation, as defined here, in mind, it suffered from a number of serious deficiencies in that regard. The main deficiencies parallel the general points raised in Chapter 5: the absence of indicators for at least some of the most important policy outcomes, the absence of comparison units, and the absence of treatment randomization. Taken together, these shortcomings present almost insuperable obstacles to an impact evaluation. One important finding of the team was that with foresight some of these deficiencies might have been fairly easily corrected and for not much additional cost. Indeed, some of the changes outlined below would likely yield cost *savings*.

As mentioned, the decentralization project sought to foster citizen participation, transparency, and accountability at the local level, with the ultimate objective of promoting "increased responsiveness of subnational elected governments to citizens." Though some of these outcomes are potentially, albeit imperfectly, measurable, indicators gathered at the local level related almost exclusively to *outputs* rather than outcomes. For example, the indicators gathered 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 CSOs) attended a training course in participatory planning and budgeting; the percentage of targeted provincial governments in which at least two CSOs exercised regular oversight of municipal government operations, as measured by participation in at least two public forums during the year; and the percentage of participating local governments that establish technical teams to assist with decentralization efforts (PRODES PMP 2007).

Such indicators are designed to monitor the implementer's 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 the main outcomes that the project was designed to affect. For purposes of evaluating impact—and even for improved project monitoring—we want to know not how many training courses there were or how many officials attended them but rather whether they led subnational elected governments to be more responsive to their citizens.⁸

⁷A description of current USAID project monitoring can be found in Chapter 2.

⁸The USAID/Peru team and local contractors were clearly aware of the distinction between measures of contractor performance and measures useful for assessing impact; this

Several indicators gathered through surveys did tap citizens' perceptions of the responsiveness of subnational 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,⁹ asked: 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? (PRODES PMP 2006, 2007).

In principle, such survey questions may provide useful proxy measures of the outcomes of interest. In practice, however, 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 all three surveys, allowing for a very limited time series on the outcome of interest. Second and perhaps more importantly, 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 socioeconomic 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 subnational 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 project. As far as the team could determine, the assessment was not repeated.

USAID/Peru's implementer was tasked with carrying out the decentralization project in all 536 districts of the seven selected regions. Once the rollout of interventions in all municipalities had been completed, no untreated municipalities remained available in the selected regions. The absence of appropriate control units (untreated municipalities) is perhaps the biggest problem for effective evaluation of the decentralization project. In addition, since rollout was completed by the second year of the program, there was little opportunity to compare outcomes in treated and untreated units in the seven regions.

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 outcome measures related to the process of supporting decentralization.

⁹The 2003 and 2005 questions were administered as a part the Democratic Indicators Monitoring Survey, whereas for 2006, data came from the Latin American Public Opinion Project.

In principle, comparisons could be made across treated municipalities in the seven selected regions and untreated municipalities outside these regions. Since the seven regions were nonrandomly selected on the basis of characteristics that almost surely covary with municipal capacity and subnational government responsiveness (e.g., high poverty rates, narcotics-related activities, past presence of the Shining Path), however, inferences drawn from such comparisons would be problematic, although not completely uninformative. 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 oversample of residents in the seven regions in which USAID works (Carrión et al 2007).¹⁰ 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).¹¹ The 2006 LAPOP national survey, had it been carried out beginning in 2003, could have established a national baseline against which the selected regions could have been measured before the program began.¹² The project implementers would then have known, for example, if as was hypothesized, satisfaction with local government, participation in local government, corruption in local government, and so forth, were more problematic in the targeted regions than in the rest of the country. Since the regions selected were poorer and more rural than the nation as a whole, covariate controls could have been introduced in an analysis-of-variance design that could have statistically forced the nation and the control groups to look more alike. Then, in each subsequent round of surveys, comparisons could have been made between the nation and the targeted regions, thereby making it possible to observe the rate of change. Had satisfaction with local government nationwide remained unchanged while the targeted areas showed increased satisfaction, project impact could have been established with a reasonable degree of confidence. Indeed, if national satisfaction had

¹⁰In addition to 1,500 respondents in the nationwide sample, an oversample of 2,100 (300 per region) was taken from the seven regions (Patricia Zárate, Instituto de Estudios Peruanos, personal communication June 2007). *Inter alia*, this survey asked respondents their opinions of the quality of local government services, as noted above.

¹¹The LAPOP instruments include questions that are comparable across 20 surveyed countries; see Seligson (2006). For useful information, the committee is grateful to Patricia Zárate, Instituto de Estudios Peruanos.

¹²Of course, the national sample would need to have had removed from it any sample segments lying in the project area in order for the national “control” group not to have been contaminated by the project inputs.

declined over the life of the project while the target areas held steady, this, too, could have been an indicator of project success. It is important to stress that since the mission was already regularly conducting national samples of public opinion, *there would have been no added data-gathering costs* in the hypothetical strategy just proposed. The only cost would have been the minimal expense of analyzing the data.

Outside the LAPOP 2006 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.

Although some data were collected in control municipalities outside the seven regions, the absence of a control group *within* the regions has serious consequences for evaluation. As just one example, many municipalities in the seven regions had been ravaged by the conflict with the Shining Path during the 1980s and 1990s. Investment and population return have picked up in some areas during the past decade, especially the past five years; at least some of this upturn must be due to the end of the war and other factors.¹³ 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. If control municipalities had been selected from the outset at random and the treatment municipalities had outperformed the controls, we would have greater confidence that the project had a positive impact.

In sum, as discussed further below, if the project had been designed to permit rigorous impact evaluation rather than monitoring, a plan for gathering data on control units would have been created as part of the initial project design. Ideally, one would have compared treated and untreated municipalities *inside* the seven regions. In the absence of untreated municipalities inside the regions, data could have been gathered on appropriately selected municipalities outside the region.¹⁴ Surveys should have included residents of untreated municipalities, and evaluations of municipal capacity (such as the GRADE study) should have included pre- and postmeasures on municipalities with which USAID/Peru's contractor was *not* assigned to work.

¹³Interviews, Ayacucho, June 27, 2007.

¹⁴However, as discussed below, without assignment, data on controls may also not help with the inferential issues mentioned in the previous paragraph.

An Alternative Evaluation Design

It is possible, looking backward, to describe an ideal randomized impact evaluation design for the decentralization project that could have been implemented in 2002. Assume that the decision to implement the decentralization project in the seven nonrandomly chosen regions was not negotiable; inferences about the effect of the intervention would then be made to the districts and provinces that comprise these regions.

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 project (e.g., training in participatory budgeting, assistance for civil society groups); 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.¹⁵ Another is that district mayors and other actors may more easily become aware of treatments in neighboring districts. For both of these reasons it would 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 could be gathered. Survey evidence on citizens' perceptions of local government responsiveness would be useful, as would information on participation in local government and 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 project—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—could then be reliably attributed to the effect of the bundle of interventions, if the difference is bigger than might reasonably arise by chance.

One feature of this design that may be perceived as disadvantageous is the fact that treated municipalities are subject to a bundle of interventions. Thus, if a difference is observed across treated and untreated groups, it may not be known which particular intervention was responsible (or most responsible) for the difference: Did training in participatory budgeting matter most? Assistance to CSOs? 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

¹⁵Some interventions also occurred at the regional level, particularly toward the end of the project, yet these interventions constitute a relatively minor part of the project.

it may not be clear exactly what aspect of treatment is responsible for differences in average outcomes across treatment and control groups. Despite this drawback, it seems preferable to design an evaluation plan that would allow USAID to know with some confidence whether a project it financed made any difference. Bundling the interventions may provide the best chance to estimate a causal effect of treatment. 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.

USAID/Peru is preparing to roll out a second five-year phase of the decentralization project, 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. The committee's understanding is that there are several possibilities for the actual implementation of the second phase of the project; 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 of partitioning 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 would not allow comparison of treated and untreated cases, it would 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 CSOs? 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 project is to reduce the number of municipalities treated, for budgetary reasons. Suppose the number of municipalities were 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 as the 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 displease authorities in control municipalities as well as USAID officials who would want to choose municipalities where they believe they have the greatest chances for success. 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.¹⁶

The preceding discussion of USAID/Peru's past and present support of decentralization projects suggests that impact evaluations could be achieved by incorporating techniques of randomized evaluation. Current monitoring efforts do not give USAID evidence about the impact of investments in local government, yet such decentralization and local government strengthening projects are a staple in the USAID DG toolbox. The good news is that the committee's field team concluded that a randomized evaluation of key aspects of the Peru decentralization project would be feasible with only modest adjustments in project design.

Supporting Multiparty Democracy in Uganda

In 2007, USAID/Uganda finalized plans for two comprehensive multiyear DG initiatives in response to the changing political dynamics in the country, especially the reintroduction of multiparty politics. One project, entitled *Linkages*, aims to strengthen democratic linkages within and among the Ugandan Parliament, selected local governments, and CSOs, building on the mission's longstanding support of legislative strengthening. The *Linkages* project is intended to "assist civil society groups, local government, and Parliament to demand transparency, accountability, and more effective leadership at both the local and national levels that will ultimately result in increased and improved essential service delivery and effective democratic representation" (USAID/Uganda 2007a). The guiding hypothesis of this project is that investments in citizen participation will drive growing demands for responsiveness and thus increase the overall quality of participation, representation, and interaction across all levels of government.

The second project, comprising a set of activities to strengthen multiparty democracy in Uganda, has the goal of "increasing democratic participation, transparency, and accountability in Uganda by supporting peaceful political competition, consensus building, and capacity building of major parties" (USAID/Uganda 2007b). This effort is driven by the hypothesis that increasing citizen participation in the development of political parties will improve the overall quality of political participation, representation, response, and interactions.

¹⁶For reasons discussed above, it may also be useful to conduct the randomization at the provincial rather than district level.

Both projects are multifaceted. They involve a wide range of interventions at different levels, from support for party development at the national and local levels, to continued legislative-strengthening activities in Parliament, to capacity-building efforts with CSOs and local governments. Recognizing the complexity of these programs, the field team worked with the mission to identify a subset of distinct interventions that would be amenable to randomized evaluation. Three specific designs are described here; additional details are included in Appendix E. As with the Peru programs discussed above, the goal of this exercise is to assess the feasibility of this approach for the programs under consideration by the Uganda mission and to highlight improvements it could afford for making causal inferences about program success. As in the Peruvian case, the evaluation designs were shaped in consultation with mission staff.

Support for CSOs

One of the core activities envisioned in the Linkages project is a capacity-building program with grants to CSOs to enable them to monitor local governments and help improve representation and service delivery at the local level (USAID/Uganda 2007a). These grants are thought to have two main impacts: (1) to develop a more robust civil society by increasing the capacity of the CSOs that are awarded the grants and (2) to improve the performance of government service delivery by increasing civic input and oversight of government officials. Whether such grants indeed have these effects is a question that can be addressed using randomized evaluation.

The best possible strategy for measuring impact would involve a large N randomized evaluation. Because a large N study would require providing grants to a large number of CSOs (more than 50) and additional monitoring and measurement, the costs are greater than what is currently envisioned for CSO grants in the Linkages project. However, this design offers substantial benefits over a small N comparison and is of general interest to USAID (especially given the results of the Gugerty and Kremer (2006) study on assistance to women's self-help CSOs described in Chapter 5).

In the proposed design, across carefully matched subcounties, large grants, small grants, and no grants would be allocated by lottery to local CSOs working on HIV/AIDS.¹⁷ One goal would be to compare the impact

¹⁷An additional benefit of focusing on HIV/AIDS is that USAID/Uganda is receiving a very large infusion of funds from the President's Emergency Plan for AIDS Relief program, so information about the effectiveness of CSOs doing HIV/AIDS service delivery would serve broader mission interests.

of large grants to CSOs (treatment group 1) versus small grants to CSOs (treatment group 2) in order to determine the impact of increases in CSO funding. Providing small grants to a second treatment group would allow USAID to assess independently the effect of greater monetary resources, while controlling for the nonmonetary effects of receiving a USAID grant (such as public recognition, special accounting requirements, and outside monitoring). Both treatment groups could then be compared to CSOs in matching subcounties where no grants are awarded (control group) to evaluate the total impact of awarding a grant.

Carefully matched groups of three subcounties would be selected purposively so that the subcounties in each group are similar along a number of dimensions that are measurable and likely to be associated with CSO capacity and government service delivery for HIV/AIDS programs. Selection criteria might include the type, size, budget, and experience of the HIV/AIDS-related CSOs already working in the subcounties, as well as the subcounties' size, urban population, wealth, voting patterns, background of key officials, location, ethnic composition, number and type of health care facilities, and infection rates. The most important criteria to ensure comparability should be determined in consultations with experts. Grouped subcounties might be next to each other, but immediate proximity is not necessary (or even desirable).¹⁸

In each subcounty one CSO working in HIV/AIDS would be selected with the aim of finding similar CSOs across three subcounties in the group. One subcounty in each group would be randomly assigned to receive a large CSO grant to monitor HIV/AIDS services in the subcounty. Another subcounty in the group would be randomly selected to receive a small CSO grant for HIV/AIDS. The remaining subcounty in the group would act as the control and receive no grant. This would be repeated for at least 50 groups, preferably more.¹⁹ It is important to ensure that (1) the large grant provides a significant increase to the existing budget of the CSOs and that the small grants do not and (2) that the CSOs spend their grants entirely on HIV/AIDS activities within the selected subcounty and that there is no contamination (sharing of resources or expertise) across subcounties. It would probably work best to select CSOs that work only in a single subcounty to prevent supplementing or siphoning of funds to the treatment sites. CSOs in both treatment groups should receive equivalent

¹⁸Instead of grouping subcounties in sets of three, it might be more feasible to use an alternative stratified sampling procedure whereby all subcounties in the sample are stratified into types according to key factors and then subcounties within each stratum are randomly assigned into each of the three categories.

¹⁹Depending on the districts chosen for Linkages, it may be possible to randomly select all the treatment and control subcounties from within the 10 districts.

technical assistance and training on how to use the grant money and how to monitor and improve service delivery.

Data would be collected before the grants are awarded, after the money is given (or at several points during the grant period), and two years after the end of the grant in order to assess both short-term and medium-term impacts. Equivalent data would be needed about CSOs and service delivery in the both the treatment and the control subcounties. To study the effect of grants and increased resources on the organizational capacity of the CSOs, data would be collected on the budget, activities, operations, and planning of the CSOs. In addition, pre- and postintervention surveys could be conducted with CSO employees, volunteers, government officials and employees, and stakeholders to evaluate changes in the activities, effectiveness, and reputation of the CSOs.

To evaluate the effectiveness of CSO grants on the delivery of government services, data could be collected on HIV/AIDS services and outcomes within each subcounty. Much of these data may already be collected by the government (such as the periodic National Service Delivery Survey conducted by the Uganda Bureau of Statistics—though perhaps USAID would need to fund an oversampling of respondents in treatment and control subcounties) or perhaps they could be collected in collaboration with other donor projects. Special attention should be given during the research design stage to determine the government activities likely to be affected by greater CSO involvement and how those activities might be accurately measured. Additional data collection could be done through surveys of service recipients or randomized checks on facilities and services. In addition, money-tracking studies of local government and government agencies could be conducted to evaluate the level of corruption in HIV/AIDS projects in the selected subcounties.²⁰

Local Government Support

One objective of USAID/Uganda's Linkages program is to increase the capacity of local governments to demand better services and representation and the accountability of local governments to their own constituents. USAID calls for actions that build the knowledge and efficacy of local government leaders, strengthen public and CSO participation, and increase local government involvement in fighting corruption (USAID/Uganda 2007a:21-22). USAID also notes that, "due to over-absorption of development programs in many district centers and severe under-absorption at the sub-district level, the Contractor should propose meth-

²⁰For more information on Public Expenditure Tracking Surveys and Quantitative Service Delivery Surveys, see Dehn and Svensson (2003).

ods of working with selected local governments and civil society groups at the sub-county levels in the identified districts” (USAID/Uganda 2007a:16). Although the specific interventions were as yet undefined, the Request for Proposals suggested working with elected and appointed leaders, traditional leaders, women, youth, constituents, and CSOs at a subcounty level. Most likely, the program will consist of a bundle of interventions rather than a single activity.

The fact that USAID plans to work with a sample of subcounties (within 10 preselected districts) makes this activity an excellent candidate for randomized evaluation. The number of subcounties within the 10 districts will almost certainly be enough to provide for a large N randomized evaluation. Therefore, in planning interventions at the subcounty level, provision would be made for the random selection of treatment and control subcounties. One approach would be to randomly select half the subcounties within the 10 districts to be in the treatment group and receive the full bundle of interventions. The remainder of the subcounties would receive no interventions and thus serve as a control group. Alternatively, subcounties could be stratified along district boundaries or other criteria, and random selection could take place within strata to facilitate equivalence on important dimensions.

It is difficult to determine the most appropriate measurement tools without a better understanding of the exact interventions and the goals of the program. Regardless of the measurement approach, equivalent data would need to be collected in the subcounties in the control group as well as those in the treatment group. Ideally, baseline data would be collected before implementation of the program and then again during and after. USAID could also investigate the possibility of contributing to ongoing data collection efforts by the government or other agencies (such as the yearly school census, the service delivery survey, the Afrobarometer public opinion survey, and public expenditure tracking surveys) in order to provide the necessary funds for oversampling in the 10 selected districts. In most cases, oversampling will be necessary to obtain data that are representative at the subcounty level.

Interparty Debates

In an effort to support multiparty democracy, USAID envisions interventions to “foster discussion and dialogue among the political parties so that difficult decisions can be achieved through compromise and negotiation before they result in conflict and stalemate” (USAID/Uganda 2007b:18). Building on successful interparty dialogues during the campaign before the 2006 presidential elections, USAID is considering sponsoring local-level political debates at the district level and below to engage

citizens in multiparty politics more effectively. In thinking about how to evaluate such activities, it is natural to ask: How does exposure to interparty debates impact citizen knowledge and attitudes about politics, voter turnout, voting outcomes, and political conflict at the local level?

Randomized evaluation offers a powerful tool for assessing the impact of interparty dialogues. Five voting precincts could be randomly selected to be in the treatment group for each of 14 different parliamentary constituencies. Remaining precincts in the 14 constituencies would make up the control group. In each of the 70 treatment precincts, interparty debates would be held between candidates for Parliament in advance of the next election. Specifically, a given group of candidates vying for a single parliamentary seat would participate in interparty debates in five different precincts within their own constituency. This would take place across 14 different groups of candidates in 14 different constituencies.

Many outcomes of interest are already collected by the electoral commission—voter registration, voter turnout, and the percent vote for each candidate. If interparty candidate debates help mobilize candidates, there should be higher registration and turnout rates in treatment precincts. One might imagine also that debates inform citizens about lesser-known candidates and thus increase the vote for nonincumbent candidates or parties. Therefore, if debates create a more informed citizenry, there should be a smaller share of the vote for incumbents in treatment precincts. If, instead, debates remind voters of the greater experience and access to largess possessed by the incumbent, the opposite effect would be evident. To gain greater power, a difference in difference estimation strategy²¹ could be used to evaluate changes from the last election in turnout and vote outcomes (assuming that the boundaries of the voting precincts are relatively stable since the last election and polling-station-level data are available for the last election). An analysis of the distance of control precincts from treatment precincts can also be performed to account for the fact that citizens in neighboring constituencies in the control group may attend or learn about debates in the treatment precincts.

To assess the impact of interparty debates on local conflict, one could also compare measures of election-day violence and intimidation gathered by DEMGroup, party observers, or outside monitors.

If resources were available to conduct surveys in treatment and control precincts, the evaluation would provide an even richer perspective on citizen knowledge, attitudes, political tolerance, and behaviors, enabling a better understanding of the causal pathways linking debates with registration, turnout, and vote choice. Ideally, pre- and posttreatment panel surveys would be carried out in treatment and control sites. Of course,

²¹See Chapter 5 for a description of this evaluation design.

care must be taken to ensure that the population surveyed in the treatment sites is comparable to those surveyed in the control sites. For example, it would be misleading to survey only those individuals who attended the debates in the treatment sites but to survey a random sample of individuals in the control group (including those who would have attended if the debate were held in their area and those who would not have). A random sample of all adult citizens in both treatment and control groups would be more informative.

While the field team in Peru described how a past project might have been designed in a way that permitted rigorous evaluation, the Uganda team focused on a multifaceted set of projects that were just getting started. Working with mission staff, the committee's experts identified a series of planned interventions, each of which could be assessed using tools of randomized evaluation. Although these evaluation models do not cover every planned intervention currently under consideration by the Uganda mission, if implemented, they would provide substantial new evidence about the efficacy of USAID DG programming in Uganda.

CHALLENGES IN APPLYING RANDOMIZED EVALUATION TO DG PROGRAMS

The evaluation designs described above are the basis for the unanimous conclusion of the field teams that randomized evaluations, apart from being valuable where they can be successfully applied, are also feasible designs for measuring the impact of (at least some) ongoing USAID DG projects. Yet demonstrating the feasibility of designing randomized evaluations that do not require significant modifications of "normal" DG projects does not imply that adopting them will not involve at least some trade-offs. Indeed, USAID staff and implementers in all three countries visited raised objections and concerns about some of the problems that randomized evaluations might pose. While several of these problems do, in fact, constitute real obstacles to program implementation or evaluation, the field teams concluded that alternatives exist in many cases that could help partially or wholly address the concerns that were raised. This section discusses these problems and how randomized evaluations could be designed to minimize them.²² Two important issues that are deferred and discussed separately—the former in the next chapter and the latter in Chapter 9—are the questions of what to do with projects that treat too few units to be suitable for randomized evaluation and problems arising from

²²See Savedoff et al (2006) for another discussion of objections to rigorous evaluations and ways they can be overcome.

the incentives (or disincentives) that DG staff and implementers have to conduct impact evaluations and their current capabilities to do so.

Randomly selecting units for treatment is simply not workable. Adopting the principle of random assignment runs the risk that certain units that project designers would very much like to include in the treatment group will wind up being excluded from the program. For some USAID staff and implementers with whom the committee spoke, this was a major reason to resist adoption of randomized evaluations. It was pointed out, for example, that in many situations USAID and its implementers can only work with local authorities that accept their help. Moreover, it was suggested that units (municipalities, ministries, groups) that lacked the “political will” to work with USAID to fully implement the programs in question would not be likely to achieve successful outcomes and thus do not merit an investment of resources. It was also suggested that units with exemplary past performance sometimes appeared to be such sure bets for program success that excluding them from participation in the new project appeared wasteful.

These are reasonable objections; however, accepting their merit need not imply jettisoning a randomized design. One option that satisfies the need for randomized selection of treatment units while also recognizing that rolling out a program in some units may not be feasible would be to select the set of units that are eligible for treatment on the basis of political will and other criteria that USAID believes maximize the chances for success and then to assign units randomly to treatment and control groups within this group of eligible units. This approach is also useful for situations where USAID seeks to limit programs to needy or conflict-affected areas, as long as there are more units than USAID can possibly treat.

Another option, suitable for situations where, for political or other reasons, allocating treatment to one or several units may be nonnegotiable (i.e., the consensus among project designers is that a particular unit or units simply *must be* included in the treatment group), is to go ahead with random selection of units for treatment but leave aside a certain percentage of the project budget (e.g., 10 to 15 percent) to pay for the implementation of program activities in units that were not selected but that organizers feel must be included. In such a case the evaluation would be based on a comparison of the regular treated group (not including the added units) with the control group. Of course, one can always look as well at outcomes in the non-randomly selected—the “must have”—units. Yet comparing outcomes in such units to nontreated units would be less informative about the causal impact of the USAID intervention than comparing outcomes across the units that were randomly assigned to the treatment group and the control group.

It is unethical or impossible to preserve a control group. Is it ethical to deny treatment to control groups? This issue arises frequently in public health programs but may also be relevant in projects where, as with interventions in the area of DG, the assistance is welfare improving even if not, strictly speaking, life saving. As with public health studies, the standard defense applies: Without an experiment, how do we know whether or not the intervention helps? USAID intervenes to assist DG all over the world. As in the public health field, it behooves us to know with as much confidence as possible what works and what does not. Continuing to channel scarce resources to projects that, once properly evaluated, turn out to have no positive impact is wasteful, particularly when properly executed randomized evaluations could put USAID in a position to identify projects that do work and whose reach and impact could usefully be expanded with a shift in resources from those that have been found to be underperforming.

A second defense of randomized assignments against the criticism that some units will go untreated is that, in any project being implemented across a large number of potential units, there will virtually always be untreated units. In the context of a decentralization project involving dozens of municipalities, it is simply not feasible for USAID to work with all of them; in the context of a project designed to support CSO development, it is simply not possible for USAID to work with every group. Given the impossibility of treating *every* unit, the only question is how untreated units will be chosen. In many contexts it may be fairest, and most ethically defensible, to choose untreated units by lottery, as would be the case in a randomized evaluation.

Finally, even if every unit is to be treated, it may be reasonable to delay treatment for a portion of the units by a randomized rollout. In this case, while some units (chosen by lottery) will get assistance first, others will have a delay before they receive assistance. Yet for the group that faces delay, this may be more than compensated by the possibility that the delayed group will either be spared an ineffective treatment or will receive improved assistance, since the initial phase of the rollout provides the basis for learning from a randomized impact study of the treatment's effects.

Isolating control from treatment groups is not feasible in practice. A third objection involves the great difficulty in preventing the effects of treated units from "spilling over" and affecting control units. For example, a project that provides support for CSOs to advocate improved service delivery may impact not only the area in which the CSOs are based but also neighboring areas (either because local governments fear similar mobilization and act to forestall it or because CSOs in neighboring areas become emboldened by the example of what their colleagues are doing

next door and step up their own advocacy). Another example of spillover is when grassroots party activities in one locale yield benefits in other places, either because party contacts extend across administrative boundaries or because changing attitudes are transmitted across familial and social networks. Whenever there are spillover effects (and there often are), the difference between the control and treatment groups is attenuated, and this will bias the evaluation toward a finding of no effect.

Sometimes, design modifications can help minimize the likelihood of spillover. For example, in the context of the Peruvian decentralization project discussed earlier, randomizing at the provincial level might decrease the probability that district mayors are aware of treatments administered to other units. In this case all municipalities in a province would be in either the treatment group or the control group, thereby minimizing the likelihood of spillover from municipality to municipality (except insofar as they happen to be located adjacent to a provincial boundary).

But while problematic for inference, spillover effects may be important to measure in their own right. In their study of deworming programs in Western Kenya, for example, Miguel and Kremer (2004) found that deworming interventions are not cost-effective *unless* the positive externalities of the program that spill over into neighboring untreated communities are accounted for. Taking advantage of the fact that the treatment is randomly assigned across space, they estimate the size of these spillover effects and then use the estimates to calculate the true effects of the deworming program, which they find to be positive once the spillover effects are accounted for. Their study underscores that not just minimizing but also measuring contamination must be a core aspect of any well-designed randomized evaluation.

A related problem is the possibility that donors from other countries might concentrate their programs in areas in which USAID is not undertaking program activities, thereby, as one program officer put it, "flooding the controls." This may happen intentionally, when donors coordinate and divide up areas of focus to avoid duplication of efforts. Or projects not intended to directly influence democracy, such as programs to create entrepreneurs or regional cooperative associations, may in fact help the spread of democracy in the area being observed. If this occurs, the other donors' interventions become a confounding factor associated with treatment, and this will almost certainly bias inferences about the effect of USAID interventions.²³

One possible response to this issue is not to advertise the existence of

²³However, it might be pointed out that, if anything, this is likely to dilute the (it is hoped positive) effect of treatment. If other donors flood the controls and there is still a difference between groups, a causal effect of USAID's intervention can be inferred. (At least, the effect of USAID relative to other donors can be evaluated.)

control units. For example, in the context of a decentralization project it may be known that USAID is working in seven regions, but it need not be made publicly known which particular municipalities it is working with in each region. A second solution is to commit in advance to implement the project in all units (and to make this publicly known) but to roll it out gradually, using untreated units as a comparison group for treated units in the years before they are added to the intervention (as in the second design for the Peru decentralization program described earlier). Another option is to randomize different treatments across all municipalities. In other words, USAID would work with *all* municipalities in the seven regions (thereby leaving no municipalities to be flooded) but randomly assign different treatments to different municipalities (again, as discussed earlier for Peru). One final possibility is to engage other donors in conceptualizing the evaluation exercise. If multiple donors are implementing similar interventions, all would benefit from an impact evaluation of their projects. In such circumstances it may be possible to coordinate USAID's activities with theirs to preserve a control group.

It is hard to plan an evaluation (or stick to one) because mission objectives and programs change all the time. A common concern the field teams heard was that randomized evaluations are insufficiently flexible to be practical. As a political officer at the U.S. Embassy in Peru commented, the embassy is sometimes compelled to “put out fires.” For example, in an experimental evaluation of the impact of municipal-level interventions in mining towns, the embassy might have to intervene if a conflict broke out in a community. This may or may not pose an issue for causal inference. Some “fires” may be independent of treatment assignment—that is, they may be equally likely to occur in treated units as in control units. However other “fires” may be products of the treatment. They may reflect, for example, the absence of a desired treatment among controls, which necessarily feel left out. This raises more serious issues. Unanticipated events that require additional interventions in either treatment or control communities must be recorded so that they can be taken into account in the final evaluation. Such events may make interpretation of the results more complicated, but the possibility that they might arise is not an argument to forego randomized evaluations per se.

In addition, missions may wish to adjust programming midstream, either by learning lessons from an early assessment of outcomes or by responding to new developments on the ground. Sometimes this is quite consistent with the purposes of a good evaluation. For example, if there is powerful evidence part way through that a project is working, USAID may wish to extend its reach into communities that were previously in the control group (medical trials are often abandoned early if there is robust

evidence of the benefits, or dangerous consequences, of a treatment). The phrase “if there is powerful evidence” is crucial here. Since the whole purpose of the randomized evaluation is to generate evidence for a project’s success or failure, there is no trade-off whatsoever in abandoning it or in tweaking it midstream, if “powerful evidence” for the project’s efficacy has already emerged. A real trade-off presents itself only if the evidence for the project’s success or failure is still tentative. In such a situation a judgment call would have to be made about the relative importance of confirming what the initial evidence seems to suggest (which would require not altering the design of the randomized evaluation) or moving ahead with the change in course (which might have the benefit of maximizing impact but risks acting on a hunch that may have been ill founded).

The more difficult issue is when, as frequently happens, unforeseen challenges arise in project implementation that USAID thinks require slight adjustments in the interventions or sometimes the replacement of implementers. Changing the treatment part of the way through the process is, of course, not ideal. As long as the adjustments are consistent across the treatment group, however, there is no threat to causal inference (although it should be kept in mind that the ultimate evaluation measures a more complicated treatment). Whatever the source of the midstream correction, responsible officials will need to remember that the benefits of continuing with the rigorous evaluation design accrue agency-wide and are not limited to the particular mission or project. So the advantages of a midcourse correction for a project or mission will need to be balanced against the potential loss of valuable evaluation information that could be usefully applied to programs in other countries.

Randomized evaluations are too complex; USAID does not have the expertise to design and oversee them. Staff both in the field and in Washington consistently raised the objection that USAID is not well equipped to design and implement, or even simply oversee, randomized evaluations. This is a valid concern. While the idea of randomized evaluation is intuitive and easy to understand, the design of high-quality randomized evaluations requires additional academic training, specialized expertise, and good instincts for research design. It is likely that many (or most) USAID DG staff do not have training in research methods and causal inference, thus making it difficult for them to evaluate the quality of proposed impact evaluations or to play a role in their design and implementation.

The committee wants to emphasize that the guidance provided in this report should not be seen as a “cookbook” of ready-made evaluation designs for DG officers. It would be a mistake for USAID to endorse the typology of evaluation designs outlined in Chapter 5 and then require DG

officers to put these new designs into practice without additional training or support. Because the issue of competence and capacity is so central to the prospect of improving evaluation in USAID DG programs, Chapter 9 is dedicated to providing recommendations about how USAID DG could make the necessary investments and provide appropriate incentives to encourage using impact evaluations of its projects where appropriate and feasible.

It will cost too much to conduct randomized evaluations. Perhaps the most important objection the committee encountered in the field is that randomized evaluation will cost too much. In part, this is a question of USAID's priorities. If the agency is committed to knowing whether important projects achieve an impact, it will need to commit the necessary resources to the task. But aside from whether the agency commits to higher quality evaluations, it is legitimate to ask how much more randomized evaluation will cost than the procedures currently employed.

The committee's field teams were tasked with some detective work in an effort to answer this question. As discussed in Chapter 8, the committee discovered that USAID could not provide concrete information about how much it spends on monitoring and evaluation (M&E) every year, even for a subsample of DG programs. The committee therefore encouraged the field teams to explore the cost of current approaches by reviewing project documents and through discussions with mission staff. They, too, encountered insurmountable obstacles; project documents almost never provided line items for M&E and what was reported was not consistent from one project to another. Based on interviews with implementers, the field teams reported that nontrivial amounts of time were dedicated to the collection of output and outcome indicators and the monitoring of performance, but no team could arrive at any hard numbers related to current expenditures. The committee thus cannot answer the question of how much more it will cost to introduce baseline measures, data collection for comparison groups, or random assignment, relative to current expenditures on M&E. At best it can be said that in a number of cases that the field teams examined, it seems that substantial improvements in all of these areas could be obtained for little or no additional cost, but that in other cases the costs could be substantial. Much depends on whether data are being collected from third parties or local governments versus being generated by surveys or other primary data collection by implementers, on whether surveys are already being used for the projects or would need to be developed specifically for the project in question, and on the specific outcomes that have to be measured in the treatment and control groups. As noted, in some cases—such as reducing the initial number of units treated in order to preserve a control group—an impact evaluation could

actually save money compared to providing all groups with assistance immediately, before the effects of the project have been tested.

But how much will a randomized evaluation cost? Answering this question requires two different calculations. The first is the straightforward calculation of how much more it will cost to collect the necessary data. This will depend on the number of control and treatment units required for a useful random assignment; the more subtle the expected effects, the larger the number of units that will be required, with a corresponding increase in the cost of data collection. The factor to keep in mind is that, even if data collection is more costly in a randomized evaluation design, the potential benefit is that it would put USAID in a position to assess the impact of the project with much more confidence and to detect subtle improvements that might not be visible without a randomized design.

The second, much trickier, calculation lies in assessing (1) the cost of selecting units at random, which may entail not implementing project activities in units where USAID might have reason to believe that the project will have a large positive impact and/or (2) going ahead with the implementation of project activities in units where USAID has reason to believe that the project will fail. Here the cost is less a direct expense than an opportunity cost. Again, these costs must be weighed against the potential benefit of being able to conclude whether or not the project worked. Note, however, that the latter type of cost (of directing program funds either to places where staff are convinced the project will not work or away from places where staff are convinced that it will) will be greater the more confident staff members are about whether or not (or where) an accurate prediction can be made about exactly where a project will be successful and where it will not. If it is already known whether (or where) a project will work, then randomized evaluations are not needed to answer this question. The real peril lies in believing wrongly that the consequences of a program are, in fact, known and allocating resources on that basis when the hypotheses behind a program have *not* been tested by impact evaluations.

CONCLUSIONS

The committee's consultants believed they had demonstrated that at least some of the types of projects USAID is now undertaking could be subject to the most powerful impact evaluation designs—large N randomized evaluations—within the normal parameters of the project design. For a majority of committee members, this provided a “proof of concept” that the designs would also be feasible in the sense that they would work in practice as well as in theory. However, one committee member

with experience in actually managing DG programs remained skeptical as to whether the complexity and dynamic nature of DG programming would allow random assignment evaluation designs to be implemented successfully. The committee also notes that doing random assignment evaluations in the highly politicized field of democracy assistance will likely be controversial. It is, therefore, recommended in Chapter 9, as part of a broader effort to improve evaluations and learning regarding DG programs at USAID, that USAID begin with a limited but high-visibility initiative to provide a test of the feasibility and value of applying impact evaluation methods to a select number of its DG projects.

REFERENCES

- Carrión, J.F., Zárate, P., and Seligson, M.A. 2007. The Political Culture of Democracy in Peru: 2006. Latin American Public Opinion Project (LAPOP), Vanderbilt University and Instituto de Estudios Peruanos, Lima, Peru. Available at http://stemason.vanderbilt.edu/files/gcflNu/Peru_English_DIMS%202006with20corrections20v5.pdf. Accessed on April 26, 2008.
- Dehn, J., and Svensson, J. 2003. Survey Tools for Assessing Performance in Service Delivery. Working Paper. Development Research Group, The World Bank, Washington, DC.
- GRADE. 2003. Grupo de Análisis para el Desarrollo, Línea de Base Rápida: Gobiernos Subnacionales e Indicadores de Desarrollo. Lima, Peru: GRADE.
- Gugerty, M.K., and Kremer, M. 2006. Outside Funding and the Dynamics of Participation in Community Associations. Background Papers. Washington, DC: World Bank. Available at <http://siteresources.worldbank.org/INTPA/Resources/Training-Materials/OutsideFunding.pdf>. Accessed on April 26, 2008.
- Miguel, E., and Kremer, M. 2004. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* 72(1):159-217.
- Olken, B.A. 2007. Monitoring Corruption: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy* 115:200-249.
- PRODES PMP. 2006. Pro Decentralization Performance Monitoring Plan, 2003-2006. Lima, Peru: ARD, Inc.
- PRODES PMP. 2007. Pro Decentralization Performance Monitoring Plan, Fifth Year Option, February 2007-February 2008. Lima, Peru: ARD, Inc.
- Savedoff, W.D., Levine, R., and Birdsall, N. 2006. *When Will We Ever Learn? Improving Lives Through Impact Evaluation*. Washington, DC: Center for Global Development.
- Seligson, M. 2006. The Americas Barometer, 2006: Background to the Study. Available at: <http://sitemason.vanderbilt.edu/lapop/americasbarometer2006eng>. Accessed on February 23, 2008.
- 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/Peru.
- USAID/Uganda. 2007a. Request for Proposals (RFP): Strengthening Democratic Linkages in Uganda. Kampala, Uganda: USAID/Uganda.
- USAID/Uganda. 2007b. Request for Proposals (RFP): Strengthening Multi-Party Democracy. Kampala, Uganda: USAID/Uganda.

Additional Impact Evaluation Designs and Essential Tools for Better Project Evaluations

INTRODUCTION

The previous chapter explored whether randomized evaluations could be more than just a theoretically appealing methodology but could also feasibly be designed for democracy and governance (DG) projects being implemented in the field by the U.S. Agency for International Development (USAID). This was done by describing a decentralization project in Peru and a series of democracy-strengthening activities in Uganda and by showing how randomized designs could be developed that would suit the implementation of these projects. Also addressed were some of the objections that the committee's field teams heard about the viability of adopting randomized evaluations more generally. While concerns about the impracticality of randomized evaluations must be taken seriously, in principle many of them could be dealt with through creative project design and/or greater flexibility in the selection of units for treatment or the timing of project rollout.

The committee recognizes, however, that randomized designs are not always possible and alternatives need to be considered. This may be because of the costs, complexity, timing, or other details of the DG project. Thus this chapter focuses on other methods of impact evaluation for those cases where randomization is not feasible. Examples are given of ways that USAID can develop sound impact evaluations simply by giving more attention to baseline, outcome, and comparison group measurements. The chapter begins by addressing two questions regarding choices between the use of randomized designs or the other (comparison-based) impact

evaluation designs described in Chapter 5. First, how many of USAID's current projects appear suitable for randomized impact designs? Second, when projects are not suitable for randomized evaluations, what options are available and how should the other methods described in Chapter 5 be chosen and applied?

HOW OFTEN ARE RANDOMIZED EVALUATIONS FEASIBLE?

To help answer this question, project staff collected information about the DG activities that the USAID mission in Uganda had undertaken in recent years (see Appendix E for a list of these projects as well as those in Albania and Peru). The projects in Uganda included efforts designed to provide support for the Ugandan Parliament, strengthen political pluralism and the electoral process, and promote political participation—a fairly typical roster of projects and one that parallels those implemented by USAID missions in many countries. A team member then divided these projects into 10 major activities and scored them for (1) amenability of each activity to randomized impact evaluation and (2) where randomized evaluation was not deemed possible, the benefits of adding other impact evaluation techniques (better baseline, outcome, or comparison group measures) to existing monitoring and evaluation (M&E) designs.¹ In doing so, the committee recognizes that current USAID project monitoring plans are largely designed to track an implementer's progress in achieving agreed-upon outputs and outcomes. Our approach, therefore, is not to assess the quality of current monitoring plans but rather to assess and illustrate instances where additional information that could reveal the impact of DG projects is currently not being collected but could readily be acquired.

The first finding of the analysis was that *all 10* of the activities examined used M&E plans that omitted collection of crucial information that would be needed if USAID sought to make impact evaluations of those activities.² The committee does not mean to criticize current M&E plans, which focus on acquiring important information for program management and resource allocation. The committee wants to draw attention to the marked difference between the content of the currently mandated and universal M&E components of most DG projects and the information that would need to be acquired to conduct a sound and credible impact evaluation of project effects. The latter is a different task and, as noted,

¹This section is based on the work of Mame-Fatou Diagne, University of California, Berkeley.

²See Chapter 2 for a discussion of the difference between current USAID project M&E plans and impact evaluations.

may require different expertise in designs for project implementation and data collection than are currently part of USAID's routine activities. For example, unless collection of data from a nontreatment comparison group is an explicit part of the project design, there is no need to monitor whether contractors are collecting such data, and it will not normally be part of M&E activities. But without such data (including good baseline data) and a set of policy-relevant outcome measures, a project's actual effects, as opposed to the accomplishment of project tasks, such as the number of judges trained or improved municipal accounting systems established, cannot be determined.

On a scale of 1 to 10, with 10 being the most complete and credible plan for collecting data for impact evaluation, 9 of the 10 activities received a score of 1 and one received a score of 2. Again, this underlines the difference between the character of currently mandated M&E designs and impact evaluation designs. Nonetheless, on the positive side, 5 of the 10 activities were found to be, in principle, amenable to using randomized evaluation designs to determine project impacts; 4 other activities were found to be amenable to collection of baseline or nonrandom comparison group data that would significantly improve USAID's ability to know whether or not the activity in question had a positive impact. Seven of the 10 activities were found to be amenable to changes in how outcomes were measured that by themselves would markedly strengthen the monitoring they were already doing.³ The measurement changes alone were judged to be capable of bringing the average ability to provide inferences about project outcomes from 1 to 3 on the 10-point scale, while the shift to collecting data for impact evaluation designs was found to be capable of raising the average score for making sound inferences of project effects to over 7. These are dramatic changes, and they underscore the team's conclusions about the large potential for USAID to more accurately and credibly assess the effects of its DG projects by adding efforts to collect impact evaluation data to its M&E designs, in at least this subset of its ongoing projects.

While the scoring of these monitoring efforts is necessarily subjective and the ability to generalize from the efforts being implemented by a single mission is obviously limited, analysis of the Uganda mission's DG activities nonetheless offers some useful lessons. First, it suggests that a number of avenues to improve knowledge of project effects are possible, ranging from simple changes in how outcomes are measured to more substantial yet feasible changes in evaluation design. Second, it suggests an answer to the question posed earlier about the frequency with which

³The team's conversations with both the mission and the implementers in Uganda included a number of discussions about the problems of measurement for DG projects.

randomized evaluations are likely to be feasible. In Uganda at least, randomized evaluation was judged to be a feasible evaluation design strategy for 5 of the 10 activities being undertaken, and an additional 4 out of the 10 were judged to be amenable to nonrandomized yet systematic baseline/control group designs. In effect, then, 9 out of 10 programs in Uganda could have potentially benefited from the approaches presented in this report. This is a much larger share than is commonly assumed by the USAID staff with whom the committee consulted in the course of its investigations.

Critics are right that randomization is often not possible, however, and the team judged that for evaluating the impact of one-half of the activities examined, only other forms of evaluation designs (i.e., the large N nonrandom comparison or small N and single-case comparisons discussed in Chapter 5) would be feasible. Yet the team's finding that one-half of the DG activities it examined were amenable to randomized design is a higher proportion than most critics would expect. This would indicate that claims that randomized impact evaluations are only "rarely" or "hardly ever" possible may be too pessimistic. Perhaps even more important, fully 9 out of 10 of these activities were found to be suitable for some form of the impact evaluation designs described in Chapter 5. Given that none of these activities in Uganda are currently collecting the kind of information needed for such impact evaluations, but 9 out of 10 could potentially do so, USAID appears to have a great deal of choice and flexibility in deciding how much, and whether, to increase the number of programs and the amount of information it collects to determine the effects of its DG activities.

As noted in Chapter 5, randomized evaluations require that there be a very large number of units across which the projects in question might, at least in principle, be implemented, as well as that program designers be able to choose these units randomly. Many high-priority USAID DG projects—for example, those that focus on strengthening individual ministries, professional associations, or institutions; those that support the creation of vital new legislation or constitutions; or those that build capacity to achieve national-level goals such as more effective election administration—do not meet these criteria. Such projects are critical to achieving the larger goal of improving democratic governance. Precisely because they are important, improving USAID's ability to evaluate the impact of the millions of dollars that it spends each year on implementing such projects should be accorded a high priority.

The next section addresses the question of what to do to carry out impact evaluations in situations of this kind. First, the general issue is discussed and then the other evaluation techniques highlighted in Chapter 5, with specific examples from the field are discussed. Finally, the discussion

takes up the special, but common, case of how to design the most credible impact evaluations when there is only a single unit of analysis.

DESIGNING IMPACT EVALUATIONS WHEN RANDOMIZATION IS NOT POSSIBLE

As stressed above, all sound and credible impact evaluation designs share three characteristics: (1) they collect reliable and valid measures of the outcome that the project is designed to affect, (2) they collect such outcome measures both before and after the project is implemented, and (3) they compare outcomes in both the units that are treated and an appropriately selected set of units that are not. As long as the number of units (N) to be treated is greater than one, all three of these attributes of impact evaluation are possible. The major difference between randomized evaluations and other methodologies lies in the degree to which project designers need to concern themselves with the number and selection of control units. In a randomized evaluation the law of large numbers does the job of ensuring that the treatment and control groups will be (within the limits of statistical significance) identical across all the factors that might affect the project impacts being measured. When random assignment is not possible, project designers must pay close attention to the factors that might be associated with inclusion in the control or treatment groups—what social scientists refer to as “selection bias”—and the effects of those factors on the differences found between the control and treatment units. These are the approaches referred to as large N and small N comparisons in Chapter 5.

Aside from the fact that the implementer does not select treated units at random, the examples described below are very similar to the randomized designs. In particular, they share the key characteristics that reliable and valid measures of project outcomes still must be collected both before and after project implementation and for treatment and comparison groups. As with randomized designs, the discussion proceeds by providing examples of best practices. All four examples highlight the importance of finding an appropriate way to identify a control group, while the latter two also emphasize creative ways to improve measurement.

National “Barometer” Surveys as a Means to Design Impact Evaluations for Localized USAID Project Interventions

For a variety of reasons, USAID often implements programs at a subnational level, applying its efforts in a selected set of municipalities or departments or regions. Often the selection of these regions is determined by programmatic considerations. For example, USAID might determine

that it wants to focus its resources on the poorest areas of the country or on areas that have suffered the most from civil conflicts or have been hit with natural disasters. In other cases USAID decides to focus on municipalities or regions that look the most promising for the success of a particular intervention. In still other cases, USAID engages with other donors to “divide up the pie” with, for example, the European Union agreeing to work in the north while USAID works in the south. Finally, there may be entirely idiosyncratic reasons for the choice of where to work (and where not to work) related to the preferences of individual host governments or implementers.

In each subnational project the principle of randomized selection is violated and the possible confounding effect of “selection bias” would be an important factor in designing an impact evaluation. The nonrandom selection may bias the impact so that, *ceteris paribus*, the results may be better than they would have been had randomization been used to select the treatment area or they could be worse. It is impossible to know beforehand exactly what to expect. The point is that those who wish to study impact will worry that selection bias by itself could be responsible for any measured “impact,” rather than the project itself.

Consider a project carried out in an exceptionally poor area. One possible outcome is that the area is so poor, and conditions so grim, that short of extraordinary investment, citizens will not really notice a difference. Similarly, in a post-civil war conflict, feelings of hatred and distrust may be so deeply ingrained that project investments will be ignored entirely. In these cases, even though the project may have been designed well, any impact is imperceptible. On the other hand, in both cases, the very low starting point suggests (as noted in the Peru example below) that a “regression to the mean” is inevitable and therefore improvements will occur with or without the project intervention. In such a case a positive impact might mistakenly be attributed to the project when, in fact, the gains are occurring for reasons entirely unrelated to the inputs.

When randomization is not possible, but selection of multiple treatment and control areas is, conditions are ideal for the “second-best” method of large N nonrandomized designs. This sort of design is often referred to as “difference in difference” (DD; Bertrand et al 2004). The objection to this approach, however, is that USAID would be spending its limited resources to study regions or groups in which it does not have projects and may not plan to have any. The committee believes that this entirely understandable (indeed compelling) reason alone constrains many DG programs and project implementers from considering a design that would be seen as “wasting” money and effort on studies in areas where USAID is not working.

The committee believes that USAID already possesses the ability to

overcome this problem of “wasting money” on seemingly irrelevant controls without significant additional investment of resources. The agency’s Latin American and Caribbean bureau, for example, is already applying this methodology in some of its projects in a limited number of instances.⁴ The approach to reduce (but not eliminate) the risks of potentially misleading conclusions is to utilize the increasingly prevalent public opinion surveys being carried out in Latin America, Africa, and Asia, collectively known as the Barometer surveys. High-quality nationally representative surveys are regularly being carried out by consortia of universities and research institutions, many with the assistance of USAID but also with the support of other donors, such as the Inter-American Development Bank, the United Nations Development Program, the European Union, and local universities in the United States and abroad. These surveys provide fairly precise and reliable estimates of the “state of democracy” at the grassroots level, by producing a wide variety of indicators. For example, the surveys reveal the frequency and nature of corruption, victimization, and the level of citizen participation in local government, civil society, and the judicial process. They also produce measures of satisfaction with institutions such as town councils, regional administrations, the national legislature, courts, and political parties and the willingness of citizens to support key democratic principles such as majority rule and tolerance for minority rights. These surveys also allow for disaggregation by factors such as gender, level of urbanization, region, and age cohort.

Given that investments are already being made in the Barometer surveys, they provide a “natural” and no-added-cost control group to studies of project impact. They provide, in effect, a picture of the “state of the nation” against which special project areas can be measured. In other words, USAID would continue to gather baseline and follow-up surveys in its project towns, municipalities, or regions and thus concentrate its limited funds on collecting detailed impact data for the places or institutions in which it is carrying out its projects. It would not need to carry out interviews of control groups for which it does not have ongoing projects. The national-level control group, however, could be used to show differences between the nation and the project areas in terms of not only poverty, degree of urbanization, and so forth but also many of the project impact measurements that USAID requires to determine project success or failure. For example, if a project goal is to increase participation of rural women in local government, comparisons could be made between the baseline and the national averages, and then, following the DD logic,

⁴The committee believes, but was unable to document, that this method has been utilized in some other programs in Africa.

comparisons would be made over time as the project impact is supposed to be occurring.

There are several recent examples to illustrate this. For many years USAID focused a considerable component of its DG projects in Guatemala on institution building at the national level, especially the legislature. Surveys carried out by the Latin American Public Opinion Project as part of its Americas Barometer studies, found a deep distrust in those institutions, despite years of effort and investment. It also found special problems in the highland indigenous areas. In part as a result of those surveys, the DG programs in Guatemala began to shift, focusing more on citizens and less on institutions. As part of that strategy, every two years national samples were carried out, along with targeted special samples (what USAID calls “oversamples”) in the highland indigenous municipalities. A finding from those surveys was the low level of political participation among some sectors of the population. In 2006 those surveys were used to focus the “get out the vote” campaign for the 2007 election, a critical one in which a former military officer was a leading candidate.

In Ecuador a series of specialized samples have been drawn in specific municipalities, with the results being systematically compared to national samples, drawn every two years since 2001. CARE, in cooperation with the International Migration Organization, has been working in a series of municipalities along the border with Colombia, a region in which the possible spread of narco-guerrilla activities could have an adverse impact on Ecuador. Thus the municipalities were not selected at random, but national-level survey data have allowed for comparison of starting levels, so that those implementing the project would have far more than anecdotal information about the level of citizen participation in and satisfaction with local government. The survey data also allow for comparisons over time to see if trends in the project areas are more favorable than in the nation as a whole. Similar efforts have taken and are taking place in Honduras, Nicaragua, Colombia, Peru, and Bolivia.

Surveys have also increasingly been used to measure the impact of anticorruption programs, in some cases by comparing “before” and “after” impacts on a specific sector (e.g., health in Colombia) and in other cases comparing the results for the nation as a whole before and after implementation of an anticorruption program (Seligson 2002, 2006; Seligson and Reccanatini 2003). The most recent survey of citizen perceptions of and experience with corruption, supported by USAID/Albania, was released while the committee’s team was in Albania (Institute for Development Research Alternatives 2007).

For this approach to be successful, national surveys, as well as specific surveys carried out in project areas, need to be at least minimally coordinated so that the questions asked in both are identical. It is well

known that small differences in question wording or scaling can substantially affect the pattern of responses. If, for example, local government participation is an impact objective of the mission, problems will arise if the national sample asks respondents whether they have attended a local government meeting and the project sample asks how many times in the past 12 months they attended a local government meeting.

There are two potential objections to this approach. The first is cost. Surveys are thought to be expensive, but often the costs appear to be larger than they really are. In many of the countries in which USAID has democratization programs, the cost of a well-administered survey can be quite reasonable.⁵

A second objection readers may have to the DD approach is that the target (or project) areas are indeed different from the national samples in many of the ways mentioned above. Often they are poorer and more rural and therefore are expected not only to begin at levels below the nation as a whole but also to perhaps exhibit slower progress. One of the strengths of this design is that such differences can be detected and noted when the baseline survey data are collected. To correct for those differences, the survey analysis can then use an analysis-of-variance design, in which the national sample becomes merely one of the groups being compared to the various treatment regions or municipalities. Covariates

⁵Costs vary directly by hourly wages in any given country. In low-wage countries, surveys can be quite inexpensive. For example, surveys in many Latin American and African countries can be conducted for \$15 to \$25 per interview (sometimes less) as an all-inclusive cost (sample design, pretests, training, fieldwork, and data entry). For a typical sample of 1,200 respondents (which would provide sample confidence intervals of ± 3.0 percent), total costs to obtain the data would be about \$30,000. Of course, that is for one round of interviews; if the typical project involves a baseline survey followed by an end-of-project survey to measure impact, those costs would double.

Gathering the data is one cost, but analysis is another. The cost of analysis depends entirely on the price of contracting with individuals qualified to analyze such data. At a minimum, such individuals should hold a master's degree in the social sciences, with several courses in statistics. Individuals with such qualifications are often available in target countries, and an extensive analysis of the data could be obtained in many for \$20,000 or even less. Unfortunately, many of the studies the committee has seen conducted for USAID limit themselves to reporting percentages and summary statistics. Analysis of that type is rarely useful, since indices of variables normally need to be created, logistic and OLS regression techniques must be applied, and reporting of significance levels and confidence intervals is required. For example, if the consultant's report states that the baseline study finds 10 percent of respondents attending municipal meetings in both the control and experimental areas, and the end-of-project survey finds that the treatment area has risen to 15 percent but the control group has also risen to 12 percent, it would be important to know if the change in the treatment group is statistically significant and if the increase in the control group was also significant. Thus USAID needs to be certain it has hired qualified individuals and obtained an appropriate level of statistical analysis to make the analysis useful for determining project impact.

can and have been used to statistically remove the impact of the differences between the national sample and the treatment groups. Hence, if the targeted areas are, on average, poorer or exhibit lower average levels of education, those variables can be included as covariates to “remove” their impact, after which the nation and the treatment areas can be more effectively compared.

There are certainly possible flaws in this sort of analysis; for example, if there are unmeasured differences that are not known and/or cannot be controlled for statistically, the findings could be deceptive. But when randomized assignment cannot be used, this method can provide a good alternative. Since in many cases missions will not be able to select their treatment areas randomly, the “national control” sample offers a reasonable way of measuring project impact.⁶

Finally, it is important to add that survey samples should *not* be used when little is known about the expected project impact. Surveys are best used when researchers already have a good idea of how to measure the expected impact. For example, in the illustration mentioned above, it should be relatively easy to specify what increased participation means, by devising questions on frequency of attendance at town meetings, municipal council meetings, district meetings, and the like. But when a project involves less well-researched areas, focus groups should be the instrument of choice until researchers more fully understand what is going on. Focus groups can then lead to more systematic evaluation via surveys.

Strengthening Parties: An Example from Peru⁷

Another example of an impact evaluation design when randomization is not possible comes from Peru, where one of USAID’s programmatic goals is to strengthen political parties. An idea that has been considered by the Peru USAID mission that would serve this goal and reinforce the parallel goal of promoting decentralization is to provide assistance to

⁶Another factor to consider with respect to the use of surveys is the size and nature of the sample size of both the treatment and the control groups. The key factor here is the change that the project is expected to make on the key variables being studied. For example, if, again, the goal of the project is to increase participation in local government, what is the target increase that has been set? If the increase is 3 percent, a sample of 500 respondents will be too small, since a sampling error of about ± 4.5 percent would emerge from a sample of that size. This means that the project evaluation would be subject to a Type II error, in which the expected impact did indeed occur, but the sample size was too small to detect it. Ideally, the control group(s) should be of the same size as the treatment group in order to maintain similar confidence intervals for the measurement of project impact/nonimpact.

⁷This discussion is drawn from the report of a field team led by Thad Dunning, Yale University.

the major national-level parties in opening or strengthening local offices. Because of the large number of municipalities in which such offices might, in principle, be opened or strengthened, such a program might seem like a good candidate for a randomized evaluation. To set up the ideal conditions for an impact evaluation, USAID or the local implementer would randomly select municipalities in which to establish or strengthen local parties from a set of acceptable municipalities. Local parties would have to accept that USAID or the contractor would select the municipalities.

However, when and where a political party chooses to open (or allocate resources to strengthen the operations of) a municipal office is purely the business of the political party. For USAID to make such decisions would be to go well beyond its mandate of supporting good governance more generally. From a project evaluation standpoint, however, the problem is that if the parties themselves choose where to open (or allocate resources to strengthen the operations of) local offices, the design would be nonrandom. If several years into the project USAID finds political parties to be stronger in the treatment municipalities, was this due to the project or to the fact that the parties selected those local branches that were already in the process of strengthening themselves? Unless the project also provided for some local branches that the national parties did not select for funding, which likely is not feasible, it would not be possible to answer this question.

Moreover, if outcomes are not tracked in municipalities in which USAID partners do *not* support local party offices (i.e., controls), any inferences may be especially misleading. Suppose measures of local party strength are taken today and again in five years and an increase is found. Is this due to the effect of party-strengthening activities supported by USAID? Or is it due to some other factor, such as a change from an electoral system with preferential voting to closed party lists, which would tend to strengthen party discipline, including, perhaps, that of local parties?⁸ With a control group of municipalities, it could be tested whether they too had experienced a growth in party strength (in which case the cause was most likely the law, which affects all municipalities in the country, not the USAID program, which was present only in some). The point is that without data on any comparison group to provide controls,

⁸Such a change is currently being considered in Peru. 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).

it will be impossible to separate the effect of USAID local activities from the effect of the law. So at a minimum, collecting data in a set of control municipalities would be highly advantageous. Thus, even if USAID gives political parties full control over which municipalities they choose for party strengthening with USAID assistance, USAID would benefit from seeking a list of those municipalities and choosing to also gather data from a sample of municipalities not on the list, to serve as a (nonrandom) comparison group.

When units cannot be randomly assigned to assistance or control groups, the challenge for an evaluator is to identify an appropriate control group—one that approximates what the treatment group would have looked like in the absence of the intervention. In this context this would mean identifying municipalities that the parties do not select that are in all other ways similar to the municipalities in which the parties elect to work. Statistical procedures—in particular, propensity score matching estimators—have been developed to assist in the process of carefully matching units to approximate a randomized design. Alternatively, evaluators can exploit the discontinuities that exist when treatment is assigned based on a unit's value on a single continuous measure. For example, if parties elected to work in the top 20 percent of municipalities in terms of their base of support, a comparison could be constructed that exploited the fact that those just above the 20 percent threshold are quite similar to those just below.

These procedures require high-quality data on the characteristics of units that were and were not selected, as well as an understanding of the factors that contributed to the selection process. But as discussed in Chapter 5, these approaches have already been employed with impressive results in other settings not too dissimilar from some DG activities. The larger point is that creativity can help overcome some of the potential obstacles to stronger research designs. And as long as they include a control group and sound pre- and postmeasurements, even nonrandomized designs can provide the basis for credible impact evaluations; in principle they can offer considerably more information for assessing project effects than is usually obtained in current DG M&E activities.

Supporting an Inclusive Political System in Uganda⁹

Another example is the project sponsored by USAID's Uganda mission to promote the development of an inclusive political system. A key objective of this effort is to empower women and other marginalized citi-

⁹This discussion and the following one draw on work by a team led by Devra Moehler, Cornell University.

zens to lobby district and political party leaders on issues of importance to them, such as activities for the disabled. To achieve this objective, small grants are to be provided to a small number of civil society organizations (CSOs) to allow them to carry out programs in this area. The objective is certainly worthy, but it is not amenable to randomized evaluation without a substantial increase in the number of funded CSOs (see Chapter 6). How, then, can it be determined whether the money spent on the small grant program is having the desired effect?

The current M&E plan for the project involves a participatory evaluation, primarily an analysis of survey data on whether respondents thought the projects “were helpful or very helpful,” supplemented by discussions with recipient organizations. A major limitation of this approach is the lack of a comparison group; data were collected only from groups or citizens who received USAID support (i.e., that were “treated”) and no effort was made to collect additional data from groups or citizens who did not receive USAID support (i.e., that could serve as a “control”). Any changes identified in the data attributed to the project might just as easily have been caused by confounding trends that happened to be taking place at the same time and that affected *all* communities (the project was implemented during an election period, so the more general effects of electoral mobilization cannot be ruled out as an alternative explanation for the observed changes in lobbying activism). Even in a small *N* design, an impact evaluation design (as opposed to the current M&E approach) that tracks trends both before and after a program is implemented and explicitly identifies untreated units for which comparable outcomes could be measured would provide much greater confidence in any inferences about the project’s actual effects.

If there are large amounts of data, the techniques described earlier (propensity score matching, regression discontinuity) can be employed. In this context, however, there is no substitute for careful, qualitatively matched comparisons. For example, if three districts were selected in which to implement the program, the evaluator would need to identify three additional districts that are similar on a set of variables believed to be associated with the targeted outcomes (e.g., income, government capacity, infrastructure). More qualitative approaches mirror the logic underlying the quantitative techniques—the goal is to identify a relevant counterfactual in order to distinguish the impact of the program from spatial or temporal trends that, while outside the ambit of the DG assistance program, could influence outcomes in the areas being observed.

The measurement strategy in the existing M&E plan could also be significantly improved. The use of subjective assessments of activities by their participants raises two concerns: (1) because they are subjective rather than objective and (2) because the satisfaction of participants (par-

ties, CSOs, etc.) is not necessarily the same thing as project success and thus cannot provide reliable information about project's impact. So one major area where improvement would be possible is providing additional external or objective measurements of program success (e.g., how much more funding for help for the disabled was actually granted to districts where CSOs received USAID assistance than was granted to otherwise comparable districts?).

Building the Capacity of the Parliament in Uganda

Another example is the case of the bundle of USAID-sponsored activities designed to build the capacity of the Ugandan parliament through the sponsorship of field visits, public dialogues, and consultative workshops for members of parliament and parliamentary committee staff regarding specific issues such as corruption, family planning/reproductive health, and people with disabilities. The project sponsored fact-finding monitoring and supervisory field visits to 35 districts, including a number in Northern Uganda, where many members of parliament and parliamentary staff rarely venture. Again, the goals of the project are worthy and the activities appear to be well conceived; however, the project is not amenable to randomized evaluation. How can it be known whether or not the money spent on project activities had any demonstrable positive effect? Did members of parliament who participated in these activities behave differently than those who did not?

As is often the case with such projects, the principal monitoring method for these activities involved the collection of quarterly data on "outputs" (i.e., the number of public meetings attended by parliamentary committee members at the local level, the number of CSOs submitting written comments to parliamentary committee hearings, etc.) rather than "outcomes" (such as the impact that workshop attendance had on information acquisition, job performance, or other aspects of future behavior). Also, the reports submitted by the implementing contractor do not provide much information on how the locations where the various public meetings took place or the participants who were invited to attend were selected—both of which are crucial for ruling out selection effects. The indicators measured by the contractor as part of the performance measurement plan of the project were used as indicators of project success.

However, because of the absence of a control group, it is impossible to disentangle time-varying unobserved trends from the impact of the project. For example, it is difficult to conclude that an increase in the number of parliamentary committees responding to CSOs with briefings and dialogues is an indication of project success. Such a change could reflect other (local) dynamics, the impact of other donor programs, the impact of the

project of interest, or a combination of these. Similarly, in the absence of a counterfactual, the fact that the Persons with Disabilities Act was passed and enacted without executive initiative or support cannot be assumed to reflect the impact of the project.

As with the projects described previously, an evaluation design that furnishes more information for assessing impact than the current M&E approach is possible. First, assessing the impact of these initiatives would require some measurement of outcomes among a control group of members of parliament who were not exposed to the field visits, public dialogues, and consultative workshops. Perhaps with the intervention defined so broadly, identifying a control group is too difficult. By focusing on a more narrow set of activities, such as the opportunity for members of parliament to participate in field visits or facilitated consultative meetings between parliamentary committees and their constituencies, envisioning a reasonable control group is more feasible. For example, if not all members of parliament are going to participate in field visits, one simply needs to understand the selection process for members (and the differences that exist between participants and nonparticipants) in order to rule out characteristics correlated with participation in the program that might account for any observed differences after the field visits (i.e., members of parliament already engaged in the conflict elect to take part in a field visit to Northern Uganda). It might be possible to facilitate a series of consultative meetings for one committee at a time and to compare how behavior changes in that committee to other similar committees that had not yet benefited from the program.

In terms of the measurement of impact, one simple improvement could involve interviewing members of parliament about their actions and opinions rather than their perceptions of the usefulness of program activities. For example, instead of (or in addition to) asking, "If you participated or were aware of these activities, how useful were they in helping to generate government action on the problems in Northern Uganda?" (the current questionnaire item), a better approach would be to ask members of parliament at the beginning and after the program about their opinions on the conflict in Northern Uganda and about what they thought should be done and any action they have taken or intend to take. Questions aimed at measuring precisely what actions, if any, members of parliament or parliamentary committees took following the field visits would provide a better sense of the effects on behavior. If these questions were asked of both participants and nonparticipants, analysis of the differences between "treatment" and "control" members of parliament would be possible. Even if these questions were asked only of participants but both before the intervention and afterward, analysis of the changes in the opinions and actions of "treatment" members of parliament would be possible.

The advantage of this type of evaluation design is that it permits analysis of changes or differences in members' actual opinions and actions rather than their subjective assessment of the "usefulness" of programs. In addition, collecting information on the basic characteristics of those members who participated and those who did not would allow some statistical matching of the two groups to better determine how much the USAID DG program, as opposed to other prior characteristics of the members, contributed to any observed differences between the two groups in their subsequent actions and opinions.

As with the two other projects described earlier, implementing the proposed changes involves trade-offs, but the team concluded that, if USAID wished to learn more about the precise effectiveness of these programs, there is substantial opportunity to develop impact evaluations on these activities, even without using randomized designs.

WHAT TO DO WHEN THERE IS ONLY ONE UNIT OF ANALYSIS¹⁰

Many USAID projects involve interventions designed to affect a single unit of analysis. Such interventions are among the most important DG-promoting activities that the agency underwrites. But for the reasons explained in Chapter 5, they are also among the most difficult to evaluate.

For example, a major part of USAID's DG-related activities in Albania involves increasing the effectiveness and fairness of legal-sector institutions. While critically important to the mission's goals, almost none of the rule-of-law activities are amenable to randomized evaluation or other methods that exploit comparisons with untreated units. This is because they each deal with (1) technical assistance to a single bureaucracy (e.g., Inspectorate of the High Council of Justice, Inspectorate of the Ministry of Justice, High Inspectorate for the Declaration and Audit of Assets, Citizens Advocacy Office, and National Chamber of Advocates); (2) support for the preparation of a particular piece of legislation (e.g., Freedom of Information Act and Administrative Procedures Code, a new conflict-of-interest law, and a new press law); or (3) support for a single activity, such as implementation of an annual corruption survey. For a randomized evaluation of the efficacy of these activities to be possible, they would have to be, in principle, able to be implemented across a large number of units, which these are not. There is only one Inspectorate of the High Council of Justice, only one conflict-of-interest law being prepared, and only one National Chamber of Advocates being supported, so it is not

¹⁰This section and the next one draw on the work of a team led by Dan Posner, University of California, Los Angeles.

possible to compare the impact of support for these activities both where they are and are not being supported and certainly not across multiple units. The best way to evaluate the success of these activities is to identify the outcomes they are designed to affect, measure the outcomes both before and after the activities have been undertaken, and compare these measures. Collecting high-quality baseline and follow-up data, the former stretching back as far in time as possible, is the primary tool for impact evaluation in such a situation. When outcome data show a marked shift subsequent to an intervention and examination of other possible events or trends shows that they did not correspond to this shift, a credible case can be made for the intervention's impact.

One problem, however, is that finding appropriate measures of the outcomes that the activities are designed to affect is frequently far from straightforward. For example, the goals of the technical assistance to the Inspectorates of the High Council of Justice and the Ministry of Justice are to improve the transparency and accountability of the judiciary and to increase public confidence in judicial integrity. The latter can be measured fairly easily using public opinion polls administered before and after the period during which technical assistance was offered and then comparing the results. However, measuring the degree to which the judiciary is transparent and accountable is much more difficult. Part of the problem stems from the fact that the true level of transparency and accountability in the judiciary can only be ascertained vis-à-vis an (unknown) set of activities that should be brought to light and an (unknown) level of malfeasance that needs to be addressed. For example, suppose that, following implementation of a program designed to support the Inspectorate of the High Council of Justice, three judges are brought up on charges of corruption. Should this be taken as a sign that the activities worked in generating greater accountability? Compared to a baseline of no prosecutions, the answer is probably yes, at least to some degree, although one would also want to know whether prosecutions were selective, based on political reasons. But knowing just how effective the activities were depends on whether there were just three corrupt judges who should have been prosecuted or whether there were, in fact, 20, in which case prosecuting the three only scratched the surface of the problem. To be sure, 3 out of 20 is better than none, so the program can be judged to have had a positive impact in at least some sense. But knowing the absolute level of effectiveness of the program may be elusive. Parallel problems affect other rule-of-law initiatives, such as efforts to improve the ability of lawyers to police themselves.

A slightly different evaluation problem arises with respect to activities designed to support the drafting of various pieces of legislation. One fairly straightforward measure of success in this area is simply whether

or not the law was actually drafted and, if so, whether it included language that will demonstrably strengthen the rule of law. But assessing whether or not USAID's support had any impact requires weighing a counterfactual question: Would the legislation have been drafted without USAID's support and what would it have looked like? If the answers to these questions are that the legislation would not have been drafted or that the language in the resulting law would not have been optimal, the support from USAID can be judged to have been successful to the extent that the result observed is better than this counterfactual outcome. The broader problem, however, is that achieving the overarching strategic objective of strengthening the rule of law will involve more than just getting legislation drafted; it will involve getting legislation passed and then having it enforced. The point—echoing a theme from Chapter 3—is that the measurable outcome of the USAID-sponsored activity is several steps removed from the true goals of the intervention, and any assessment of “success” in these areas must be interpreted in this light. Proper measurement of project impact must move beyond proximate questions (were the institutions created?) to more distant and policy-relevant ones (have the outcomes that the existence of the new institutions were hypothesized to affect been altered in a positive way?). Answering the second question requires the existence of high-quality baseline data, preferably stretching back as far in time as possible so as to be able to distinguish general trends from project effects.

Additional Techniques to Aid Project Evaluation When $N = 1$

In addition to collecting high-quality baseline and follow-up data, two other techniques can aid project evaluators in making sound judgments about project efficacy. The first is to explicitly attempt to identify and rule out alternative explanations. If what looks like a project effect is identified, evaluators must ask what other factors outside the scope of the project might have caused the observed outcome. Can they be ruled out? For example, suppose it is found that the passage of a new anticorruption law whose drafting was sponsored by USAID corresponds with a drop in corruption, as measured in national surveys. It would be important to think carefully about other factors that might have occurred at the same time as passage of the new legislation which might also account for the drop in measured corruption. Perhaps a crusading anticorruption minister was appointed right after the new legislation was passed. Might her presence at the helm of a key ministry have caused the change? One way to rule out this possibility would be to see whether larger changes in perceived corruption were evidenced in her ministry than in others or whether perceived corruption increased again after she left office—both

of which would be consistent with the argument that her appointment, not the new law, was responsible for the drop in corruption measured in the surveys. The more such competing explanations can be identified and ruled out, the more confidence there can be in the conclusion that the legislation was responsible for the positive outcome.

Evaluators are in a better position to rule out alternative explanations to the extent that USAID or its implementing partners can manipulate the timing of the intervention. An effort can be made even before a program is begun to identify other planned interventions or major events that could affect the outcome of interest and make it hard to disentangle the effect of USAID's program from other possible factors. In this context a decision could be made to delay or speed up implementation of the program to minimize the likelihood that temporal changes in the measurement of program outcomes reflect things other than USAID's program. To make this idea more concrete, imagine an intervention designed to increase the quantity and quality of debate in a parliament. The intervention might involve a series of training sessions on parliamentary business, a change in the rules that ties salary to attendance and participation, or an accountability mechanism that reports to the public on the activities of members of parliament. Regardless of the intervention, the outcome of interest is clear: whether members exhibit higher attendance rates and are more active in parliament after the project is complete. The problem is that many other factors might be responsible for an increase in attendance or participation—for example, if preparations for the budget begin soon after the program is initiated, this may drive up attendance and participation. If these other factors can be anticipated and avoided in planning the timing of the intervention, even stronger inferences can be drawn from temporal trends in the outcome variables.

A second strategy for improving causal inference in an $N = 1$ design is to look beyond the narrow outcome that the project was designed to affect and try to identify other outcomes that would be consistent with positive project impact. The example provided earlier from Uganda of using the success of projects targeting the disabled to verify the effectiveness of completely separate projects designed to promote the empowerment of marginalized citizens illustrates this technique. With regard to evaluating the effectiveness of the anticorruption legislation, an example of such a strategy would be to look at changes in applications for business licenses, which might be expected to rise as the requirement that applicants pay bribes diminishes. Again, the greater the number of outcomes consistent with project success that can be identified, the more confidence there can be in inferring that the project was, in fact, successful.

Designing impact evaluations where a large number of units are available and USAID has control over where or with whom it will work is

relatively straightforward, although the actual design requires substantial skill. In principle, all that is needed is a random number generator—or even just a coin to toss—to assign units to treatment or control groups. Then once the project is implemented, all that is needed is to compare average outcomes in the control and treatment groups and test whether the differences are statistically significant. The higher art of impact evaluation comes in situations where randomized evaluations are not possible. Under such circumstances, identifying sound project designs requires flexibility, creativity, understanding of the facts on the ground, and a good sense of the implications of various design decisions for the interpretation of program evaluations. This makes them difficult, both to design and, because of the need to tailor the methodology to the details of the particular project in question, to specify *ex ante*. However, it does not make them impossible. As the many examples provided in this chapter suggest, there are opportunities to move beyond the current M&E approach to impact evaluations that provide key information for determining program effects, even in the most difficult, and quite common, situation where there is only a single unit being treated. Good designs require skilled, well-trained program designers—the cultivation of which should be a priority for USAID. It also requires an organization with the resources and capacity to do the work—issues discussed in Chapters 8 and 9.

CONCLUSIONS

For every DG-promoting activity that USAID undertakes, particularly those that are central to its mission or that involve the expenditure of large sums of money, USAID wants to be able to answer two questions: Was doing the activity better than doing nothing at all? If so, how much better? Generally, although they may serve other management purposes well, the required M&E designs that USAID currently employs are insufficient to do this. Answering these questions requires the use of impact evaluations, which in turn require somewhat different designs. The committee found that the vast majority of USAID staff that it encountered were deeply committed to improving democratic governance around the world and to being able to evaluate the progress they were, or were not, making. The committee also found that many USAID staffers were frustrated by their inability to better answer the basic question: Are we having a positive impact?

The impact evaluation designs described in this report, and the examples presented in the previous two chapters, suggest that in principle there is considerable scope for USAID to improve its ability to answer this question. The committee would neither expect nor recommend that the agency undertake impact evaluations of all of its activities. The com-

mittee's specific recommendation is that USAID begin with a modest and focused initiative to examine the feasibility of applying such impact evaluation designs, including those using randomized assignment, to a small number of projects.

At the same time, the committee realizes that undertaking more impact evaluations alone will not provide the broadly based and context-sensitive information that USAID needs to plan its DG programs. Process evaluations, the kinds of case studies discussed in Chapter 4, and more informal lessons from the field obtained by DG staff, implementers, nongovernmental organizations, and independent researchers provide important insights, valuable hypotheses, and illustrations of how programs are received and respond to changing conditions. The committee believes that USAID needs to develop organizational characteristics that will provide both incentives for more varied evaluations of its projects and mechanisms to help agency staff absorb, discuss, and continually learn from a variety of sources about those factors that affect the impact of DG programs.

REFERENCES

- Bertrand, M., Duflo E., and Mullainathan, S. 2004. How Much Should We Trust Difference-in-Difference Estimates? *Quarterly Journal of Economics* 119(1):249-275.
- Institute for Development Research Alternatives. 2007. *Corruption in Albania: Perception and Experience: Survey 2007, Summary of Findings*. Tirana: Institute for Development Research Alternatives and Casals & Associates.
- Seligson, M.A. 2002. The Impact of Corruption on Regime Legitimacy: A Comparative Study of Four Latin American Countries. *Journal of Politics* 64:408-433.
- Seligson, M.A. 2006. The Measurement and Impact of Corruption Victimization: Survey Evidence from Latin America. *World Development* 34(2):381-404.
- Seligson, M.A., and Recanatini, F. 2003. Governance and Corruption. Pp. 411-443 in *Ecuador: An Economic and Social Agenda in the New Millennium*, V. Fretes-Cibils, M.M. Giugale, and J.R. López-Cálix, eds. Washington, DC: World Bank.

Creating the Conditions for Conducting High-Quality Evaluations of Democracy Assistance Programs and Enhancing Organizational Learning

INTRODUCTION

Chapter 6 addressed some of the real and perceived obstacles to carrying out impact evaluations of democracy and governance (DG) projects and discussed ways that they could, in principle, be overcome. But a much more general problem exists: organizational conditions that discourage staff from the U.S. Agency for International Development (USAID) and implementers from undertaking high-quality impact evaluations. Reviewing agency policies and practices with the goal of reducing barriers to and strengthening incentives for conducting sound impact evaluations is essential. Just as important, USAID must create and nurture the capacity to learn what works and what does not by sharing information and experiences widely and openly. This chapter first addresses the specific issues of improving organizational capacity for impact evaluations and then turns to the more general problem of creating the conditions for organizational learning.

ISSUES IN OBTAINING HIGH-QUALITY IMPACT EVALUATIONS

Any changes made to the general guidance for monitoring and evaluation (M&E) of DG projects will be carried out in the field in over 80 country missions by hundreds of implementing partners. Even with the centralization of program and budget decision making undertaken in the Foreign Assistance Reforms of 2006 (USAID 2006), USAID remains a

highly decentralized agency, and country missions have substantial discretion in how they implement and manage their programs.

The committee also recognizes that the USAID contracting process is already dauntingly complex and time-consuming, demanding much of the time that DG officers spend to develop and manage their projects. The committee thus is cautious about recommending specific solutions for the contracting of evaluations, especially as contract and procurement processes are not an area in which the committee has any special expertise. What follows is instead intended as a set of principles, drawn from research and field studies, that the committee believes will help USAID in obtaining sound impact evaluations of DG projects. Examples are offered of possible approaches to the problem, but the actual design and implementation of any changes would rest with USAID. Knowing how difficult the problems of changing contract management practices are with the current reality of USAID programming, the DG evaluation initiative recommended in the next chapter could be an opportunity to try out different approaches.

Incentive Issues

A key problem, not unique to DG or USAID, is the question of providing incentives to DG staff and implementers to undertake and complete sound and credible impact evaluations. The DG officers and implementers the committee and its field teams met shared a strong desire to be successful in promoting democracy. They are drawn to their work because they believe that democracy is a better form of government and that foreign assistance can help bring about democratic development. The problem, however, is *how* to promote democracy. From the outset, DG officers and implementers alike recognized that “doing democracy” was going to be much more difficult than other areas in development such as health and agriculture where causal relationships are better understood and impacts easier to measure. There may be formidable barriers to good policy and implementation in these other areas, but at least there is greater consensus about the basic questions of theory and measurement.

The uncertainty about fundamental aspects of DG reinforces the normal human and bureaucratic incentives to avoid documented failure, a problem that has been cited as affecting evaluations across USAID and not simply DG (Clapp-Wincek and Blue 2001, Savedoff et al 2006). In the absence of a strong learning culture that encourages open reflection and recognizes the uncertainties surrounding DG programming, carrying out projects that produce no effect (or a negative effect) could understandably be considered a threat to a USAID officer’s career. Similarly, program implementers worry about their organizations’ futures and the results of

being associated with a documented failure, knowing that it is generally not the way to win future contracts or grants. In the democracy promotion area, where there is little hard evidence about what works and why and where many crucial factors that might make for success or failure are beyond the control of DG officers and their implementers, there is a natural tendency to confine measurements of success to those things over which one has some hope of control, such as project outputs and very proximate outcomes.

In addition, a host of time and resource pressures generally lead implementers not to take time before program rollout to gather extensive baseline data or to conserve precious resources for actual DG program support by keeping evaluation costs to a minimum (or, as the committee discovered, sometimes using funds from the M&E budget to support programming in the later stages of a project when resources grew tight). The clear priority for getting programs started as quickly as possible, and doing as much as possible with limited budgets, necessarily leads to a far lower priority for impact evaluation procedures, as these generally require some time and effort spent on collecting baseline data and data from comparison or control groups. Without strong incentives to complete sound impact evaluations on at least some DG programs and some rewards for doing so, these pressures make it highly unlikely that such evaluations will be designed into DG programs. **One task of the DG evaluation initiative recommended in the next chapter should be to address these issues and explore how to ease the task of undertaking impact evaluations within USAID's contracting and program procedures. The initiative should also examine incentives for both DG officers and DG project implementers to carry out sound impact evaluations of selected DG projects.**

Coordination Issues Regarding Strategic Assessments

USAID already undertakes a fairly time-consuming process of baseline assessment as part of its development of strategic objectives (see Chapter 2). At present, however, the strategic assessments guide policy planning (including choice of DG projects), which then result in calls for proposals. Evaluations enter later, if at all, in a way quite separate from the initial assessment process.

It would be far more productive for good impact evaluation if the strategic assessments also sought to identify which projects (if any) should be targeted for impact evaluations to determine their effects. Then any baseline information collected as part of the assessments could be designed, and made available, to support the desired impact evaluation. For example, any national or regional surveys, or interviews with possible

or intended participants, could be usefully incorporated into subsequent evaluations. Perhaps even more important, **the strategic assessment process must identify critical hypotheses guiding the planned democracy assistance program (e.g., that increasing local mobilization or nongovernmental organizations (NGOs) will reduce corruption), so that they can be clearly specified and designated for impact evaluations in the calls for proposals, if such evaluations are desired.**

Contracting Issues

The committee's research and field visits also found that the current process of awarding contracts and grants actually works against conducting impact evaluations in a number of specific ways:

- DG officers are chosen for expertise in democracy assistance and aid delivery, not for expertise in evaluation designs. Thus DG officers often felt they lacked expertise among their mission staff to prescribe or judge what would be an effective, high-quality impact evaluation design.
- Implementers, who often believed they had the expertise to undertake a richer variety of M&E activities, including impact evaluations, thought that USAID gave priority to doing the proposed work rather than M&E, and especially if budgets were tight, ambitious M&E plans would work against them in bidding for projects.
- Systematic communication among DG officers and between DG officers and implementers is limited, so there is little opportunity to share experiences and compare, and perhaps correct, perceptions of each other's expectations.
- Given the multiple steps in the contracting/grant-making process, there are many points at which decisions can be made that restrict or eliminate the opportunity to design impact evaluations into projects from the outset or not to carry them out fully once a project has begun.
- On the positive side, the basic system for program monitoring and use of indicators in place through the Automated Directives System is a good foundation, even if current practice could be improved (USAID ADS 2007). Thus the data collection required for impact evaluations seems practical if the incentives and contract procedures motivate implementers to schedule baseline, outcome, and comparison group measurements as part of the contracted DG activity.

Changes to the Contracting Process to Provide for Impact Evaluations

As already discussed, perhaps the key difference between the current approach of commissioning process evaluations when a mission sees

the need, as a separate contract issued after a project has begun or been completed or when a shift in strategy is contemplated, and commissioning an impact evaluation is that an impact evaluation needs to be treated as an integral part of a project's implementation design. Unless baseline measurements are part of the contract schedule and data collection on an appropriate comparison or control group is provided for at project inception, it is difficult—often impossible—to go back and obtain such information once a project has begun or been completed. This means that if a mission wants to obtain sound evidence of the impact of a particular project, staff will need to think about planning an impact evaluation before they have even drawn up the call for proposals for that project and make a suitable design for impact evaluation part of the original action and budget plan for that project.

Call for Proposals

When a USAID mission undertakes a new project or the next phase of a continuing one, in most cases there is a formal request for bids, called a Request for Proposals (RFP) for a contract and a Request for Applications (RFA) for a grant or cooperative agreement.¹ One required component for those responding to an RFP or RFA is a description of how the project would be monitored and evaluated. Given the strict federal rules governing competitive procurement policies, the RFP/RFA is the primary source of information available to a would-be implementer about the mission's goals for the project and requirements for a successful bidder, including M&E.

In current practice there is seldom any indication that an evaluation process is expected beyond the required Performance Monitoring Plans, which generally focus on tracking the project's activities and immediate outputs. In addition, as the committee learned, DG officers differ in how much detailed guidance they want to provide in an RFP or RFA, sometimes preferring to give the implementers, who have substantive expertise and experience, flexibility to provide most of the details of how they think the project and M&E should be carried out.

To undertake impact evaluations, RFPs and RFAs would need to contain explicit language indicating that on this occasion such an evaluation is expected. The solicitation would not need to specify the evaluation design in detail; the committee and the field teams were told that implementers would readily understand the implications of language that

¹A key distinction among the types of agreements is the amount of control that USAID has over how the award is implemented. USAID has the most control over contracts, less with cooperative agreements, and the least with grants, which give implementers wide discretion over how to carry out projects, including M&E.

called for sound impact evaluation as requiring the collection of baseline data, treatment and control groups where possible, and alternatives when the project involved an “N = 1” intervention. But the process would need to begin at this stage.

If a more detailed statement is considered preferable, a recent RFP in one of the missions that the field teams visited provides an example.² As part of the performance monitoring plan called for in the RFP for the Democratic Linkages project in Uganda, bidders were told they should have “a clearly developed strategy for assessing the impact of the program at all three levels [national, district, and subcounty] by evaluating outcomes over time (comparing pre-intervention and post-intervention values on impact variables) or by comparing outcomes in districts selected to receive the program and those that do not (matched to ensure their comparability)” (USAID/Uganda 2007:27).

Points for Impact Evaluations

Once USAID receives proposals, the bids must be evaluated. Another part of the competitive process is awarding points, which are specified in the RFP or RFA, to various parts of a proposal. One of the impediments to encouraging investment in evaluations is that relatively few points are assigned to the M&E plan and often the M&E plan is included as a subset of some other category rather than being graded on its own. The committee did not undertake an extensive examination of this issue, but meetings with DG officers and implementers and the field visits suggest that it would be rare for an M&E plan to count for much more than 10 out of 100 possible points for the overall proposal. By contrast, the experience and quality of the implementer’s chief of party might earn 30 to 40 points because management ability is considered so critical to project success.

The committee is not recommending a specific number of points for evaluation, but it does seem likely that some change would be needed to give a more rigorous evaluation plan a competitive advantage. Instead of changing the number of points, another approach would be to treat the M&E plan as a separate category, so that a high score might be a tipping point or a genuine competitive advantage. The DG office could consult with other areas in USAID, such as health or agriculture, where impact evaluations may be more common practice, for guidance on how to structure the points or process used in evaluating proposals.

²Again, as far as the committee was able to determine, these requirements were exceptions to standard practice.

Time Pressures

One of the most precious commodities once an award is made is time. As noted above, once an award is made, there is often great pressure to “move the money” as soon as it becomes available, to “hit the ground running” and “show early success.”

In principle, implementers generally have 30 to 60 days after an award is signed to develop an M&E plan for approval by the mission, which usually includes collection of some kind of baseline information or data prior to, or very soon after, the project (assistance) activities begin. Yet in practice, two things often happen: (1) time pressures mean that project activities actually begin before all the work to set up and implement the monitoring plan and baseline measurements can be accomplished or (2) the process of approving the monitoring plan can drag on, sometimes for months, so that projects fall behind schedule and plans to collect baseline measures are delayed or dropped. The effect is the same in both cases: Crucial baseline data are not collected and may not be able to be reconstructed later in the project. The opportunity for a rigorous assessment of project impact may be effectively lost.

For those select projects for which DG officers want sound impact evaluations, contracting schedules for implementers need to allow for the implementation of an appropriate evaluation design, including establishing an appropriate control or comparison group and setting up and completing baseline measurements on both the assistance and the control groups.³ Policymakers may need to be reminded that rushing to roll out projects without allowing for careful examination of initial conditions and creation of comparison groups undermines the only way to accumulate knowledge on whether those DG projects are working as intended and those expenditures are worthwhile.

Keeping Project Evaluation Independent

Ideally, the individuals or contractors who implement a project should not be the only ones involved in evaluating its outcomes. After all, they have every incentive to show success. Independent evaluations by a separate contractor that show project success are therefore much more convincing.

³Where the comparison group is part of a population already being surveyed and the baseline data can be obtained from the survey, the need to establish relationships with the comparison group is obviated. But for activities involving smaller and identifiable control groups—such as sets of legislators or judges or NGOs or specific villages that will not receive assistance in the initial phase of the program—time to establish such relationships to allow proper data collection is essential to any sound impact evaluation.

USAID has already recognized this principle in its practices for process evaluations by requiring that they be carried out by agents other than the program implementers. Yet this is easier for process evaluations, which can be undertaken after a project has begun or been completed, than for impact evaluations, which generally require that plans for data gathering and analysis be “built in” to the project in the design stage.

Once an award is given, USAID could then give separate contracts, or independent tasks within the same contract, to implementers A and B, the former to carry out the program and the latter to carry out the evaluation portion. This would leave the evaluation partner, who is receiving separate payment and rating from USAID on the quality of its evaluation, with incentives to provide the highest-quality evaluations for USAID. To minimize the risk of collusion, USAID may have to require contractors who implement a large number of projects for USAID DG offices to work with several different evaluation partners; similarly, evaluation contractors should be required to partner with several different implementers over time in order to ensure continued independence of project and evaluation agents.

Resource Issues

One of the major objections to impact evaluations that the committee and its field teams encountered is that they “cost too much.” The collection of high-quality baseline data and indicators, especially since it must be done for both those who receive the DG support and a control group that does not, can be costly, although Chapters 6 and 7 discuss ways in which at least some of those costs could be reduced. But unfortunately there is no way to analyze that objection relative to current M&E spending because USAID is not able to provide reliable estimates of those costs. This is true both for USAID Washington and for the three missions visited by the committee’s field teams.

There are several reasons that USAID cannot provide an estimate of its M&E expenditures. One reason is that there is no consistent methodology for budgeting project evaluations, so that both missions and implementers may count the same things in different ways. Perhaps more important, as already discussed there are many kinds of M&E, and the costs of some are much easier to estimate than others. The list below was developed with the assistance of USAID/Washington staff and the work of the three field teams.

- *M&E plans for each grant/contract.* As discussed above, these are required of USAID grantees and contractors and approved by USAID. Proposals/applications will typically include an illustrative M&E plan,

but these differ in the level of detail, and the cost of preparing them would be difficult to measure. Sometimes a proposal includes an estimate of costs directly related to M&E (e.g., if the implementer anticipates doing an opinion survey), but this does not always happen and is not a requirement. It is uncertain whether a project's M&E budget would include the time that staff members spend collecting data on indicators and preparing required reports. In some cases, local staff will collect the information, which is then sent to the implementer's headquarters for analysis and preparation of the required reports for USAID. In this case the costs would more likely be considered part of the project's overhead than part of the M&E costs. So project budgets might show a zero (even with a good M&E plan) or might show tens of thousands of dollars if, for example, annual opinion surveys are planned.

- *Mission Performance Management Plan (PMP)*. Required of each mission as part of meeting Government Performance and Results Act (GPRA) requirements, these set out "strategic objectives" and "intermediate results" with corresponding results indicators. Many missions will spend money to have consultants train mission staff in developing PMPs and/or help develop them. Missions might also spend money to collect some data for them. But in many cases they rely on data collected by partners or from third-party sources (e.g., the host government, local NGOs) and rely on mission staff to develop the plans and compile data and thus would not have a budget line item dedicated to PMPs.

- *USAID annual report and common indicators*. Missions were required to answer certain common questions each year for the annual report (in addition to the PMPs). Starting in FY2007, this was replaced by the common indicators for USAID and the State Department developed as part of the foreign assistance "F Process" reforms. These costs are unlikely to be included in mission budgets.

- *Self-evaluations by implementers*. Some grants and contracts include plans for the implementer to conduct its own evaluation, at the midway point and/or the end of the project. Typically these will include budgets for \$10,000 to \$20,000 to bring in people (e.g., from the home office) to do the evaluation. These may or may not include a budget to collect baseline and subsequent data.

- *Outside evaluation of grants/contracts*. These are typically requested and paid for by a mission, often when it is thought a project is not performing well or a major project is close to completion and an evaluation is part of planning a follow-on project. Again, this type of evaluation almost always consists of a team of two to four consultants who spend two to three weeks in-country and base their findings largely on interviews with a range of people (mission staff, partner staff, direct and indirect beneficiaries, local experts, and so forth). This type of evaluation costs between

\$40,000 and \$100,000, depending on the number of consultants and the amount of time spent in the country. A mission might undertake zero to three evaluations of DG projects per year, depending on a number of factors (e.g., the number of activities in the DG portfolio, whether a new strategy is due, if a major event occurs in the country, new mission staff arrive).

- *Strategic objectives final evaluations.* Missions are required to conduct a final evaluation whenever they close out activities in one of their strategic objectives. These are conducted in much the same way that outside evaluations of grants/contracts are conducted, but with more emphasis on overall impact on a sector rather than exclusively focusing on the performance of the implementers. The cost would be about the same as the outside evaluations and depend on similar factors.

With 100 overseas missions, each with dozens of projects under way at any given time, it seems reasonable to conclude that millions of dollars is spent each year on M&E, broadly defined. As discussed, impact evaluations of project effects are one component of the broader M&E task, and it would not be simply a matter of transferring funds spent on one part of the M&E function to a different task. But if some of the current approaches to assessing project impact do not, in fact, provide genuine evidence of success or failure, it would seem that there are resources that could be more productively applied, even if no firm dollar amount can be provided for them. More generally, a serious examination of the balance of effort and resources among various types of evaluation, in particular that devoted to monitoring (outcome evaluation) relative to other forms that can inform strategic decisions and assessments of program impact, could be another part of the evaluation initiative recommended in the next chapter.

IMPROVING ORGANIZATIONAL LEARNING

The results of sound impact evaluations have value for USAID only when they become readily accessible knowledge for USAID officers and that knowledge feeds into learning processes that inform policy and planning. This section looks at what happens to the results of evaluations and other data after they are obtained.

Archiving Survey Data to Build “Collective Memory”

As discussed earlier in this report, USAID makes significant use of surveys in its DG programming. The committee believes that more could be done to fully exploit the utility of surveys in the measurement of DG program impact and to support greater learning across the organization.

One finding from interviews in Washington and the field is that, more often than not, raw survey data, the basis on which key comparisons within and across countries could be made, are lost. USAID currently has no central repository for the survey data its implementers collect. Given that with only the rarest of exceptions survey data by definition are computerized and almost always stored in common formats (typically SPSS, Excel, STATA, or SAS) for which interchangeability programs (e.g., Stat-Transfer) are readily available, the labor costs and storage space requirements would be trivial. **The committee recommends, as an initial step, that the DG office develop a simple system to establish and maintain such an archive.** To emphasize how basic the tasks are, the design could be created by a library information sciences graduate student working as an intern and then maintained by a junior administrative staff person.

Archiving the data, however, is far less of a problem than being sure that all of the data end up in Washington. Other studies of general USAID evaluation practices (Clapp-Wincek and Blue 2001) and the committee's own DG-focused research found that despite requirements to do so, reports written by consultants and research organizations are not routinely sent to USAID Washington. For many years the Center for Development Information and Evaluation (CDIE) played the role of archivist for USAID. But even when CDIE was functioning, reporting was not systematic. Now that CDIE has been absorbed into the office of the new director of foreign assistance in the State Department, it is not clear how well the "collective memory" of USAID will continue to grow.

Ensuring that survey data are retained would probably require an executive decision at the bureau level or higher to impose an absolute contractual requirement that the data generated would be deposited with USAID Washington. The committee recognizes that the barriers to doing so are real, as many of USAID's DG programs are carried out by consulting firms whose contractual clauses broadly prohibit the use of their data beyond the confines of the company. Finding ways to address these proprietary issues will be essential to supporting the learning culture this committee believes USAID needs to acquire.

Using Surveys More Systematically to Build a Global Knowledge Base

To develop comparable data that can be regularly updated across the range of countries in which USAID operates, more attention needs to be paid to the systematic use of its survey data. The committee notes at the outset that the field of scientific survey research has been undergoing incremental refinement since its first use in the 1940s. Genuinely representative samples can be designed and survey data obtained at relatively modest cost, and questionnaires can be crafted that provide reliable and

valid measurement of citizens' attitudes and behaviors. In practice, most USAID missions commission surveys in an ad hoc fashion that, coupled with the lack of agency-wide coordination of survey research methodology, data collection, and data analysis, means that USAID is not taking full advantage of the prospect for greater ability to develop comparability across surveys taken in many parts of the world.

As discussed in Chapter 7, surveys can be used in one form of impact evaluation design when randomization is not possible. Surveys also provide a powerful tool to test democratization hypotheses. Does corruption erode support for democracy? Do certain ethnic groups express more intolerance than others, participate less in civil society, or participate more in protest demonstrations? These are all important questions that can be asked of the Democracy Barometers surveys, and the answers can help target and adjust DG projects.

Surveys can be used to track project success over time. To refer again to civil society participation, if USAID establishes as a project goal increased participation in a given region or among females, then repeated surveys over time can help determine the extent to which those efforts have been successful. Comparisons within a country provide important information about project impact. But to obtain data that would allow for a more general comparative assessment of democratic values and practices, surveys from multiple countries are needed. USAID needs this comparative information to be able to make a determination of how advanced or hindered democratic behaviors and practices are in any given country. For example, if it finds that corruption victimization affects 10 percent of the adult population in a given country in a single year, it needs to place these data alongside survey data obtained for other countries in order to determine if the 10 percent level is high, medium, or low.

As already mentioned, consortia of researchers around the world have been developing regional surveys of democratic values and behaviors. The earliest systematic surveys of entire regions emerged in Europe with the development of the Eurobarometer and since 2001 the emergence of the European Social Survey, which now covers 25 nations in the broadened European community. Other regions of the world also are covered by such surveys, including Eastern Europe, now included in the Eurobarometer; the New Europe Democracies Barometer, which covers much of the former Soviet Union and is currently based at the University of Aberdeen; the Asian Barometer, currently based at the National Taiwan University; and, most recently, the Arab Barometer, currently based at Princeton University and the University of Michigan.⁴

⁴Recent studies by several of these democracy barometers can be found in the July 2007 and January 2008 issues of the *Journal of Democracy*.

To the committee's knowledge, USAID has invested in two regional surveys: (1) the AfroBarometer, organized by Michigan State University and the Institute for Democracy in South Africa; and (2) the Americas Barometer, organized by the Latin American Public Opinion Project of Vanderbilt University and its partner university and think tanks in Latin America, led by the University of Costa Rica.

The committee believes that greater international coordination among existing surveys should be sought and supported. At present, even among the regional barometer surveys that USAID is partially funding, there is no central coordination across these two regions. Moreover, there are many countries in Africa in which the AfroBarometer does not operate, even though USAID does work there. At this time there is no assurance that the same core items will be asked in each region and country within Africa, nor is there any reason to believe that identical questions will be asked across regions. **The committee recommends that USAID facilitate this sort of coordination among those regional surveys it is currently funding and also explore how it might promote such coordination with the Asian and Arab barometers.** For example, a small conference could be held in Washington for the senior directors of these regional barometers to see if such coordination would be possible from administrative and financial points of view. It is obvious that within a region or country many items need to be unique to tap into the particularities of that region or country's structure. Yet there is almost certainly a common core of items that could be asked that would work universally or nearly so.

Increasing Active Learning

In addition to acquiring and storing information to shed light on DG program outcomes, another essential part of the committee's recommendations is for USAID to increase its activities for actively sharing and discussing that information. The internal and external USAID Web sites and those of individual missions provide substantial amounts of information about DG projects and often furnish links to evaluations and efforts to derive "lessons learned." Unfortunately, as with survey data, although all evaluations are supposed to be provided to the Development Experience Clearinghouse (DEC) and available on the Web, in practice a substantial fraction never make it out of implementer or mission files.⁵ In the absence of resources to pursue compliance with the requirement—and perhaps enforce some sanction for failure—the competing pressures of other tasks will mean that reporting remains a low priority. The committee believes

⁵The DEC Web site is <http://dec.usaid.gov/> (accessed on August 4, 2007). An assessment of how many evaluations reach the DEC is available in Clapp-Wincek and Blue (2001).

that the results of the evaluations undertaken during the evaluation initiative recommended in the next chapter would have to be much more readily available to have the desired effect on future USAID programming. **The committee thus recommends that transmitting reports for DEC should be an important part of each project under the proposed evaluation initiative. More generally, as part of the initiative the resources of DEC should be augmented to help ensure that all project evaluation reports reach DEC so that they can be openly available.**

The Internet offers remarkable access and opportunities, but to learn from experience, DG officers and implementers also need opportunities to meet and discuss their experiences on a regular basis. Starting in the mid-1990s, when a reorganization moved technical specialists from the regional bureaus to new “centers,” including a democracy center, annual meetings of DG officers from around the world were held with implementers in the form of “partners conferences,” which provided such opportunities. The meetings frequently included outside experts to supplement and support the learning process. CDIE also organized a series of programs that provided opportunities for USAID officers back in the United States on leave to be exposed to the latest evaluations emerging from the center. Topics generally reflected the annual USAID evaluation agenda.

A number of factors, including tight budgets for operating expenses and criticism of “extraneous” travel, have curtailed these events and a significant opportunity is being lost. **The committee believes that increasing USAID’s capacity to learn what works and what does not should include provisions for regular face-to-face interactions among DG officers, implementers, and outside experts to discuss recent findings, both from the agency’s own evaluations of all kinds and studies by other donors, think tanks, and academics.** Videoconferencing and other advanced technologies can be an important supplement, but personal contact and discussion would be extremely important to share experiences of success and failure as the evaluation initiative goes forward. This includes lessons about the effectiveness of DG projects and about successes and failures in implementing impact evaluations.

This type of meeting is especially important for ensuring that the varied insights derived from impact and process evaluations, academic studies, and examinations of democracy assistance undertaken by independent researchers, NGOs, think tanks, and other donors are absorbed, discussed, and drawn into USAID DG planning and implementation. While only USAID has the ability to develop and carry out rigorous evaluations of its projects’ impacts, many organizations are carrying out studies of various aspects of democracy assistance, and USAID’s staff can benefit from the wide range of insights, hypotheses, and lessons learned being generated by the broader community involved with democracy promotion.

While it will take some time for USAID to learn from undertaking the pilot impact evaluations, it will gain immediately from augmenting its overall learning activities and increasing opportunities for DG staff to actively engage with current research and studies on democratization. Several committee members wish to emphasize the considerable value to policymakers and DG officers of the many books, articles, and reports prepared in recent years by academic researchers, think tanks, and practitioners. Whatever the methodological flaws of these case studies and process evaluations from a rigorous social sciences perspective, this expanding literature has provided important lessons and insights for crafting effective DG programs.

Turning Individual Experience into Organizational Experience: Voices from the Field

Realizing that its DG officers often had valuable insights and experiences gained from years of implementing projects in various conditions around the world, USAID's Democracy Office began a pilot project under its Strategic and Operational Research Agenda (SORA) in 2005 to collect this information systematically. Called collectively "Voices from the Field," this pilot project attempted to use extensive anonymous interviews with DG officers who had served in two or more missions around the world to understand whether there were attributes that commonly led to project success and/or failure. In this pilot phase of the project, SORA developed a standard set of interview questions for each of its initial participants. Given SORA's mission, these questions were largely designed to elicit descriptions of the best and worst projects in which the DG officer had participated (see the interview protocol in Appendix F). It then conducted interviews with eight participants, each of which lasted about two hours. The results of these interviews revealed a wide range of responses, although common trends in project success and failure also seemed to emerge.

As part of its efforts to explore methodologies that could be used to learn from past experiences, USAID asked the committee to offer suggestions as to how the Voices from the Field project might be expanded and integrated into the overall SORA research design. Based on discussions with current and former DG officers, the committee decided to explore various options for expanding this project during at least one of its field visits (see Appendix E). Practical issues the committee wanted to understand about a potential Voices from the Field project included how frequently such interviews or debriefs should occur, who should conduct such interviews or collect such insights and experiences, and in which format(s) should the information be collected and disseminated. In addition, one issue that had not been explored in the initial pilot phase of the

project conducted by USAID was whether or not those people who work for USAID DG missions around the world as foreign service nationals (or non-American citizens) would be able to provide additional sources of insight.

While in Peru the field team attempted to address these questions through a series of meetings with current DG officers and foreign service nationals, including a dedicated meeting with two foreign service nationals with considerable DG experience. As their tenure at the missions tends to be much longer than that of career DG officers, who move from one mission to another every one to four years, foreign service nationals tend to have a great deal of institutional knowledge and experience, often in particular subfields of DG programming such as decentralization or political party strengthening.⁶ It is their historical knowledge that often provides the continuity across projects over the long term.

With regard to the frequency with which interviews or debriefings should occur, it seemed that a systematic inquiry of this sort would optimally be conducted every 12 to 18 months. This time frame would be consistent with other annual reporting requirements and would largely be reflective of the natural life span of projects that DG officers and foreign service nationals oversee. Careful timing of interviews and debriefs is an important consideration given the workload of those in DG missions.

During the initial pilot phase of the Voices from the Field project, the interviews were conducted by USAID and the transcripts of the interviews were made available to USAID, although the interviewees' names were not attached to the transcript. The committee was also interested to learn whether participants would feel more comfortable responding to an interviewer who did not work for USAID even if their responses were anonymous. There was a question as to whether or not participants would feel comfortable honestly responding when asked to identify the primary attributes of both successful and unsuccessful projects if USAID were asking the questions. During the field visit inquiries the team found that this was not a great concern to potential participants. In fact, they said they felt very comfortable providing honest responses, even when discussing less successful aspects of programs. Further, they remarked that such honest discussions were a routine part of their work at that mission. The one aspect of their work, however, that those interviewed would like to highlight to a greater extent was success in more routine matters. They expressed the desire to have a voice in sharing smaller everyday successes, which are often overlooked by bigger projects, programs, and efforts.

Finally, if these interviews were undertaken on a larger scale in the

⁶The field teams in Albania and Uganda met equally experienced foreign service nationals.

future, the committee would be interested in learning which formats may be most beneficial in both collecting and disseminating information gathered from these interviews and debriefings. In Peru, foreign service nationals in particular expressed a willingness to participate in face-to-face interviews, to complete written surveys, or to complete surveys or interviews conducted through other means such as a Web-based interface. Their primary request, however, was that the results of the interviews or debriefings be widely shared with them and with other DG professionals around the world. They expressed concern that opportunities for learning may be lost if the interviews are given and no information on the insights or lessons learned was to reach those working in the missions. There was great interest in learning from their experiences as well as those of colleagues around the world; therefore they hoped that information from such programs would flow both in from and out to the field missions.

Depending on the interview design, information collected through a *Voices from the Field* project focused on systematic debriefings of DG officers, and foreign service nationals could offer very detailed information on project implementation or more general insights about potential sources of project success or failure. These would not be substitutes for the empirical evidence that impact evaluations could offer. They could, however, complement the face-to-face interactions of annual DG officers and partners recommended above by compiling a systematic record of experience; the results of these interviews might become part of the renewed conferences, further encouraging the sharing of experiences and collective learning.

As an opportunity for continued learning from its wealth of experiences, the concept of “*Voices from the Field*” is consistent with SORA’s overall goal of better understanding what has worked, why, and under what conditions. Other organizations, such as the military, employ such systematic debriefing techniques, often with great benefit. On a more ambitious level, other, more academic uses of oral history could complement or be a resource for the retrospective studies discussed in Chapter 4. Even more ambitious efforts to use “truth telling” conferences to add information and explore the varying perceptions of key historical events that have influenced how USAID views its ability to affect democratization could potentially yield valuable insights.⁷

Given the potential benefit of learning from the insights and expertise of DG officers and FSNs, the pilot project seems to offer USAID an opportunity to gain unique project-specific information it cannot acquire through other means. If incorporated into a larger framework designed to

⁷An example from the foreign policy field is the work of James Blight and his colleagues on the Cuban Missile crisis (Blight and Welch 1989), which eventually included senior U.S., Soviet, and Cuban officials who had taken part in the decision-making process.

increase learning across the organization, “Voices from the Field” would complement other systematic approaches to gathering and employing more rigorously obtained information. **The Committee therefore recommends that USAID consider a modest investment in continuing an improved “Voices from the Field” project, the results of which would be made available to USAID DG officers and FSNs.** During the period of the evaluation initiative that we recommend in the next chapter, special attention might be given to interviews with those carrying out the new procedures for impact evaluations. If SORA decides to undertake additional retrospective efforts, either by commissioning its own case studies or systematically mining current academic research, then more ambitious oral history or “truth telling” conferences might be part of the mix.

While there is an opportunity to learn from this project, learning will only occur if that information is systematically collected and disseminated to those who may gain from that information, such as DG officers, FSNs, and other USAID employees involved in project direction and management. Further, as was clear from the discussions held in the field with DG professionals, their willingness to continue to participate in such efforts was largely linked to their ability to learn from the results. The insights and experiences collected must not only be studied, analyzed, and incorporated into a larger framework of learning, but they must also be shared in an easily accessible format with those who stand to directly gain from this information. This could be accomplished through the development of a Web-based interface where respondents could complete surveys and interviews via their work computers and also access the results of other respondents. Other dissemination options should also be considered, such as providing annual results at conferences and gatherings of DG officers and professionals. Whatever the mechanism for collection and dissemination selected, if USAID chose to continue this project, it should follow standard best practices and the results should be made widely available.

CONCLUSIONS

The potential changes to current USAID policy and practices discussed in this chapter range from specific suggestions for the contracting process to a broad shift in the organization toward a much more systematic effort to share and learn from its own work and that of others. In the next chapter we introduce a set of specific recommendations based around a DG evaluation initiative intended to increase the capacity of USAID to support and undertake a variety of well-designed impact evaluations, and to improve its organizational learning. We believe this initiative will demonstrate the value of increasing USAID’s ability to assess exactly what its DG programs accomplish, and provide guidance to help USAID

better determine which projects to use, in which conditions, to best assist democratic progress.

REFERENCES

- Blight, J.G., and Welch, D.A. 1989. *On the Brink: Americans and Soviets Reexamine the Cuban Missile Crisis*. New York: Hill and Wang.
- Clapp-Wincek, C., and Blue, R. 2001. *Evaluation of Recent USAID Evaluation Experience*. Washington, DC: Center for Development Information and Evaluation, USAID.
- Savedoff, W.D., Levine, R., and Birdsall, N. 2006. *When Will We Ever Learn? Improving Lives Through Impact Evaluation*. Washington, DC: Center for Global Development.
- USAID (U.S. Agency for International Development). 2006. U.S. Foreign Assistance Reform. Available at: http://www.usaid.gov/about_usaid/dfa/. Accessed on August 2, 2007.
- USAID ADS. 2007. Available at: <http://www.usaid.gov/policy/ads/200/>. Accessed on August 2, 2007.
- USAID/Uganda. 2007. Request for Proposals (RFP): Strengthening Democratic Linkages in Uganda. Kampala, Uganda: USAID/Uganda.

An Evaluation Initiative to Support Learning the Impact of USAID's Democracy and Governance Programs

INTRODUCTION

Nearly two decades after the U.S. government and other donors began making major investments in promoting democracy and governance (DG) abroad, a number of international studies found that surprisingly little hard empirical evidence exists about the impact of these investments (see Chapter 2 for a discussion of these studies). New cross-national quantitative research suggests that DG funding *on average* has spurred democracy, but this analysis reveals nothing about the efficacy of specific projects or activities—such as local government capacity building, investments in civil society organizations, or judicial training—that have come to dominate the U.S. Agency for International Development (USAID) DG menu (Al-Momani 2003; Finkel et al 2007, 2008; Kalyvitis and Vlachaki 2007; Azpuru et al 2008). Decades of monitoring and process evaluation reports have yielded significant amounts of data on outputs (e.g., local governments supported, nongovernmental organizations (NGOs) funded, judges trained) and valuable reflections on the process of delivering DG assistance. But as discussed in earlier chapters, they have so far provided little evidence that meets accepted standards of impact evaluation about whether these projects have strengthened local governments, contributed to more robust civil societies, or helped create more legitimate judicial sectors in the countries in which they have been implemented.

Five years from now, the committee hopes that the USAID will be in a position not only to clearly and persuasively identify the effects of its DG programs but also to claim leadership in the procedures for conducting

sound impact evaluations of them where feasible and appropriate. To do this, USAID must invest in creating an ethos of evaluation, so that at least some of its DG projects are seen as presenting valuable opportunities to learn about what works and what does not in encouraging the growth of democratic institutions and values around the world.

Earlier chapters analyzed current USAID approaches to assessment and evaluation and proposed ways to provide the evidence of project impact that USAID needs both for its own programming and for presenting and defending its programs to the broader policy community in Washington and internationally. Earlier chapters focused on the specific policy and process changes that the committee believes are needed to help USAID overcome concerns that hinder undertaking sound impact evaluations and to augment USAID's overall learning to support DG programming. This chapter outlines a suggested strategy for USAID and its Strategic and Operational Research Agenda (SORA) to implement such changes.

The committee recommends a special initiative—a synthesis of many of its earlier proposals for what USAID should do in the future—to examine the feasibility of applying the most rigorous impact evaluation methods to DG projects. Recognizing both the current skepticism in the DG assistance community about impact evaluations and the significant organizational barriers that their implementation faces given current U.S. contracting and management practices, the committee's recommendation is relatively modest, more in the way of undertaking a pilot or set of demonstration projects within the current USAID structure.

PROVIDING LEADERSHIP AND STRATEGIC VISION

Obtaining more impact evaluations to determine the effects of DG programs is chiefly a matter of setting priorities, and that is the domain of leadership. Strong leadership is essential if USAID is to become an organization that prizes learning about the successes and failures of its DG projects, whether launched in the missions, regional bureaus, or the central DG office. Because DG programs are such an important—and often controversial—part of U.S. foreign policy, the committee recommends that leadership should come from the top—in the form of a DG evaluation initiative led by a senior USAID official. **This initiative should be guided by a policy statement outlining the strategic role of investments in impact evaluations of DG programming.** It is particularly important that the “vision” behind impact evaluations make clear that gaining knowledge of what works and what does not work is the primary goal. Impact evaluations should thus be targeted as far as possible to study projects as designed and carried out; the discussion in Chapters 6 and 7

shows that actual projects—not just artificial or deceptively simple versions of them—could likely be given sound impact evaluations, including the most effective randomized designs. In addition, missions and implementers with generally good records will be positively recognized, and not sanctioned, if they uncover sound evidence that programs do not work or work poorly.

This statement would provide a valuable opportunity to adjust the balance of motivations that currently drive monitoring and evaluation (M&E) in DG. The administrator should see the need for this initiative, both to ensure the sound and effective use of the considerable increases in budgetary resources going into DG programs in the past five years and to create a leading edge for revitalizing evaluation across the agency.¹

The initiative would begin a conscious and deliberate effort to undertake the highest-quality impact evaluations (including randomized designs where possible), in order to restore a better balance among different types of M&E activities, which are now largely focused on tracking project outputs or very proximate outcomes. Impact evaluations would help USAID accumulate knowledge that would (1) distinguish project models that work from those that do not, (2) identify the conditions under which particular approaches are more or less effective, and (3) help USAID avoid costly investments that may cause harm or may simply be ineffective.

The committee's charge is limited to recommendations for improving USAID's ability to evaluate its DC projects but there could be advantages to making this an agency-wide initiative. USAID implements social programs in many parts of the agency, so the changes the committee recommends could yield much wider benefits. As discussed in Chapter 2, the World Bank has taken this approach through its Development Impact Evaluation (DIME) Initiative and NGOs such as the Poverty Action Lab at the Massachusetts Institute of Technology and the Evaluation Gap Working Group of the Center for Global Development are working to promote impact evaluations for a range of social programs.² This is a time when many policymakers, both within and outside the United States, are calling for reinvigoration and rethinking of foreign assistance programs (among myriad sources, see, e.g., Lancaster 2000, 2006; National Endow-

¹A 2006 study from the National Research Council addressed the broader issues of the decline in evaluation capacity across USAID (NRC 2006).

²Information about the evaluation gap initiative can be found at http://www.cgdev.org/section/initiatives/_active/evalgap. Accessed August 27, 2007. Information about the Abdul Latif Jameel Poverty Action Lab can be found at <http://www.povertyactionlab.com/>. Accessed on August 3, 2007. Information about the DIME initiative can be found at <http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/0,,contentMDK:20381417~menuPK:773951~pagePK:64165401~piPK:64165026~theSitePK:469372,00.html>. Accessed on August 3, 2007.

ment for Democracy 2006; Epstein et al 2007; HELP Commission 2007; Hyman 2008).

In addition to its program benefits, a DG evaluation initiative could place USAID among those in the forefront of improving development policy. Although there are sound reasons to think that impact evaluations may often not prove feasible, and committee member Larry Garber has often noted such concerns, the potential gains to accurate and defensible knowledge where such evaluations do prove feasible would be considerable. USAID is unique among donors in the range of assistance projects and the number of countries in which it operates at any given time. **The committee is thus unanimous in recommending that USAID undertake a pilot program to learn whether impact evaluations will yield new insights into the effectiveness of DG projects.**

IMPLEMENTING THE VISION: THE EVALUATION INITIATIVE

Improving the evaluation of DG programs should embrace a multi-tiered approach. Not all projects need be, or should be, chosen for the most intensive evaluation using the techniques of randomized assignment to treatment and control groups outlined in Chapter 5. Neither USAID staff nor their implementing partners currently have the capacity to implement impact evaluations widely, and these skills require time and experience to develop. Moreover, as already discussed, the skepticism the committee encountered about whether impact evaluations were feasible persuaded members that a well-organized piloting of impact evaluations on a few select programs would be the best way to start. Moving too quickly or too sweepingly could impose an unacceptably high cost on USAID's efforts to assist the development of democracy and good governance throughout the world.

Tasks for the DG Evaluation Initiative

The committee strongly recommends that, to accelerate the building of a solid core of knowledge regarding project effectiveness, the DG evaluation initiative should immediately develop and undertake a number of well-designed impact evaluations that test the efficacy of key project models or core development hypotheses that guide USAID DG assistance. A portion of these evaluations should use randomized designs, as these are the most accurate and credible means of ascertaining program impact. By key models, the committee refers to projects that (1) are implemented in a similar form across multiple countries and (2) receive substantial funding (examples include projects to support local government, civil society, judicial training). By core hypotheses the

committee refers to the assumptions guiding USAID project design that, whether drawn from experience or prevailing ideas about how democracy is developed and sustained, have not been tested as empirical propositions. Examples include the assumption that public service delivery improves if citizens have oversight over the spending of public monies or the idea that exposure to democratic practices increases people's faith in democratic institutions.

The DG evaluation initiative should identify three or four program models that are widely used in DG promotion and two or three core hypotheses that guide DG thinking on democracy assistance and then plan and conduct impact evaluations of these models/hypotheses across a range of countries or contexts over the next five years. As many of these as possible should be chosen to offer feasible designs for random assignment evaluations. However, for important programs for which USAID desires impact evaluations but for which randomization is not feasible, carefully developed alternative designs, of the types discussed in Chapter 5, should be developed and implemented.

At the end of this five-year period, USAID would have:

- practical experience in implementing the evaluation designs that can indicate where such approaches are feasible, what the major obstacles are to wider implementation, and whether and how these obstacles can be overcome;
- where the evaluations prove feasible, a solid empirical foundation to begin (1) assessing the validity of some of the key assumptions that underlie DG projects and (2) learning which commonly used projects work and which do not in achieving program goals; and
- the basis for judging how widely to apply such impact evaluations to DG program evaluations in the future.

For the majority of USAID DG projects, however, the goal should be more modest. Where USAID mission directors request them, the initiative should provide support and advice to help the missions request and oversee good-quality impact evaluations that pay attention to all three elements of good evaluation practices: a focus on *outcomes*, good *baseline* measurements, and *comparison* with untreated groups. Evaluations should include pre- and postintervention outcome measures, along with, where possible, an analysis of outcomes in a relevant control group. As Chapters 5, 6, and 7 demonstrated, a wide variety of evaluation designs aside from randomized assignments are available to help USAID accumulate systematic evidence of the efficacy of particular approaches in order to guide its decision making as new investments are planned.

To assist in the effort, the committee recommends that the USAID administrator consider establishing a social sciences advisory group for the agency. This group could play a useful role in advising on the design of the evaluation initiative, helping work through issues that arise during implementation, and developing a peer review process for assessing the evaluations undertaken during the initiative.

Resources

The five-year DG evaluation initiative should be supported with special, dedicated resources outside the usual project structure. Supporting the initiative with special funds would be another signal of the strong commitment to change. The committee is not able to provide an estimate of the likely cost of the initiative, in part because the difficulties in obtaining estimates of what USAID currently spends on M&E provide no basis for comparison. Some of the essential components are discussed below to provide a rough basis for making an estimate. But the important point is that the funds not come out of current mission program budgets that are already stretched thin.

It is also important that the resources be used to support both the special impact evaluations chosen as the chief task of the DG evaluation initiative *and* efforts by country missions to improve their evaluations or conduct their own impact evaluations on chosen projects. **The initiative should thus make its resources and expertise available to mission directors who want its support in conducting impact evaluations or otherwise changing their mix of M&E activities, in order to make the initiative an asset to the DG officers in the field rather than an additional unfunded burden.**

Capacity

One of the biggest challenges facing the initiative relates to capacity. Over the past four decades, the structure of USAID has been transformed, moving away from an in-house professional staff of development experts with a significant and substantive role in projects toward an arrangement in which development projects are prioritized, solicited, approved, and overseen by USAID officers, but projects are largely designed, carried out, and evaluated by outside implementers (NRC 2006). This shift has led to an increasing focus on time-consuming issues of grant and contract management rather than project design and evaluation. This long-term shift has taken place in parallel with the more recent changes in agency policy described in Chapter 2 toward an increased emphasis on project monitoring and the use of evaluations to respond periodically to management needs, rather than systematically assess project impacts.

One consequence of these changes in policy and in the responsibilities of USAID staff has been the erosion of the skill base and expertise required to design and oversee impact evaluations for a variety of programs and contexts. The DG officers the committee encountered were experienced and knowledgeable in substantive matters, but even if they had training in general social sciences research methods, they rarely had training or experience with impact evaluation design. The evaluation capacity of USAID's DG programs, like other capabilities, has thus increasingly shifted to the implementers who design and carry out the projects. Although the committee found in its own field visits that DG officers were, in general, quite willing to work with the committee's consultants who were evaluation experts and that the DG officers were open to considering new approaches to testing the efficacy of their programs, few of the officers thought they were capable of judging and overseeing varied impact evaluation designs without additional assistance and resources.

The expertise needed among USAID professionals and, in particular, DG officers to support the initiative deserves particular attention. USAID's past hiring in the DG field has stressed bringing in individuals with practical or theoretical training in democratic processes and institutions. This will continue to be the main area for DG expertise, but it is clearly distinct from, and not sufficient for, providing expertise in the full range of project evaluation strategies. The World Bank, health care agencies, and other foreign assistance organizations regularly hire Ph.D.-level researchers whose advanced training focused on carrying out experimental and statistical evaluation analyses to support their subject matter experts. To increase its in-house capacity to support improved evaluations, USAID will need to hire more individuals with Ph.D.s in the social sciences whose training was strong in techniques of experimental and statistical analysis that can be applied to DG projects. **The committee recommends that USAID acquire sufficient internal expertise in this area to both guide an initiative at USAID headquarters and provide advice and support to field missions as a key element of the initiative.**

The DG office, like other parts of USAID, has made use of short-term appointments to augment its expertise. In the committee's judgment, however, if the recommended evaluation initiative is accepted, the practice of having an occasional Ph.D.-trained experimental analyst as a fellow in the DG office can be helpful but will probably not be sufficient. As discussed further below, valuable assistance could be provided by outside experts through USAID's various contracting mechanisms, but the leadership and confidence that come with in-house knowledge will be important to the success of the proposed initiative.

For many years the lack of staff capacity was offset by an active agency-wide centralized evaluation office (as in most bilateral and multi-

lateral development agencies)—the Center for Development Information and Evaluation (CDIE). The DG office in particular was the subject of many detailed CDIE evaluations, including substantial comparative studies of DG projects (see, e.g., Blair and Hanson 1994, de Zeeuw and Kumar 2006). As already discussed, these were generally process evaluations and not formal impact evaluations, but they did provide systematic research intended to gather lessons and compare experiences. With the increased emphasis on project monitoring, however, CDIE had grown gradually weaker in recent years and was recently absorbed into the office of the new director of foreign assistance in the State Department.

Whether or not an independent central evaluation office should be restored is beyond this committee's charge, but the committee believes the capacity of USAID headquarters to provide significant resources and expertise to assist DG officers in the field (and perhaps other USAID programs as well) who wish to develop impact evaluations of their programs would be a valuable augmentation of USAID's in-house resources.

Partnerships to Add Capacity from Outside USAID

While the committee believes that a substantial augmentation of USAID's internal capacity for evaluation design is necessary for the proposed evaluation initiative to be effective, there is no reason that USAID's efforts to improve evaluation must be purely an in-house affair. The need for supplemental outside capacity is particularly acute with regard to impact evaluations and broad-based learning. There is no need to keep on staff sufficient experts on evaluation design to provide all the assistance requested by country missions in that regard, if USAID can find other means to deliver the required technical support to field staff at critical moments of project design, implementation, and evaluation. And many of USAID's organizational learning activities can and should be enriched by partnerships with academic institutions and think tanks exploring similar issues.

USAID has a number of options through its current contracting mechanisms to acquire this support. The committee's discussions in Washington and during its field visits suggest that a significant number of implementers already have or could readily add the necessary expertise in impact evaluations; the problem has been a lack of demand for impact evaluations as parts of calls for proposals, rather than a lack of capacity among implementers.³ As discussed earlier, the committee believes that

³Local grantees, such as NGOs, pose a different problem. Although it was found in the field visits that many local partners understood the concepts of improved outcome measures and impact evaluations, few had the training and capacity to implement new practices without assistance.

it is important to maintain independence between those implementing a program and those responsible for its evaluation, but this could be achieved in a number of ways.

Universities also offer a major source of expertise related to high-quality impact evaluations. Many university-based scholars already serve as consultants to USAID implementers on a range of DG issues. Increasingly, scholars are also partnering directly with international development agencies and NGOs to design and undertake systematic program evaluations. Mechanisms such as the Democracy Fellows program allow USAID to bring scholars onto its staff for short-term appointments.

Moreover, there is ample precedent in USAID for drawing on the expertise and resources of universities rather than individual scholars. Over several decades USAID established itself as a pioneer in research leading to development in the field of agriculture. The agency accomplished this through a wide array of partnerships (usually constructed in the form of “cooperative agreements”) with U.S. land grant colleges and universities. These were institutions that had long been carrying out the research needed to achieve better agricultural outcomes. Land grant officials were accustomed to working with state agricultural extension services, for example, providing them with technical support to detect, diagnose, and cure outbreaks of diseases and infestations threatening crops and livestock. The research was not limited to agricultural production itself but dealt with a wide range of issues, including rural credit, in which Ohio State University played a key role, or land tenure, in which the Land Tenure Center at the University of Wisconsin became the world leader. Those partnerships expanded beyond the borders of the United States into international networks of research centers dedicated to agricultural research and extension. A prime illustration is Zamorano, in Honduras, but there are many others.

When USAID embarked on democracy programs as a major effort distinct from its other programs, it did not make a comparable investment in basic research partnerships with universities to provide additional knowledge and intellectual capacity. In most cases the focus was and remains on *doing democracy* rather than studying *how to do democracy*. There were and are important exceptions, and in addition some universities are major implementers of USAID DG programs, such as SUNY Albany’s long-term efforts at legislative strengthening, or the work of the IRIS Center at the University of Maryland on issues related to economic development and governance.⁴

Although not necessary for the initial DG evaluation initiative, for the longer term USAID might consider investing resources to develop a

⁴Further information about the IRIS program may be found at <http://www.iris.umd.edu/> and about SUNY Albany’s Center for Legislative Development at <http://www.albany.edu/clcd/>.

set of agency-university partnerships designed to facilitate high-quality evaluations and research in particular sectors or issues. These partners should also be involved in designing and implementing a range of discussion/learning activities for DG officers in regard to evaluations and other research on democracy. Possible models include the “centers of excellence” funded by the U.S. Department of Homeland Security or the National Institutes of Health. In addition to providing expertise to advise programming and research to advance knowledge, such agency-university centers could assist DG—and USAID more broadly—in developing a standardized training module on evaluation techniques for DG program staff.

AGENDA FOR USAID AND SORA

As part of its charge from USAID, the committee was asked to recommend a “refined and clear overall research and analytic design that integrates the various research projects under SORA into a coherent whole in order to produce valid and useful findings and recommendations for democracy program improvements.”⁵ Various parts of this design have been dealt with in depth in earlier chapters and will not be repeated here. But the committee does want to summarize the essential elements it believes could enable SORA to continue to serve as a major resource for USAID in studying the effectiveness of its programs and providing knowledge to guide policy planning.

Retrospective Studies

SORA began its work by exploring how USAID might mine the wealth of its experience with DG programs around the world to inform its future work. Based on the study by Bollen et al (2005) and its own investigations, the committee found that the records and evaluations of past USAID DG projects could not provide the requisite baseline, outcome, and comparison group data needed to do retrospective impact evaluations of those projects. Therefore the committee recommends that the most useful retrospective studies that USAID could support, if it chooses to, would be long-term comparative case studies that examine the role of democracy assistance in a variety of trajectories and contexts of democratic development. A diverse and theoretically structured set

⁵As discussed in Chapter 1, in 2000 the Office of Democracy and Governance in the Bureau for Democracy, Conflict, and Humanitarian Assistance created SORA, which consists of a number of research activities. SORA’s goal is to improve the quality of U.S. government DG programs and strategic approaches by (1) analyzing the effectiveness and impact of USAID DG programs since their inception and (2) developing specific findings and recommendations useful to democracy practitioners and policymakers.

of case studies could provide insights into overall patterns of democratization that could improve strategic assessment and planning (see Chapter 4). If USAID chooses first to take advantage of current research in the academic and policy communities, it could undertake an effort to engage systematically with those producing research and serve as a vital bridge to accumulate and disseminate evidence and findings in the most policy-friendly format possible. If USAID chooses to support case study research of its own, the committee has suggested some key characteristics for a successful research design.

Strategic Assessment

Chapter 3 made the case for a significant effort by USAID, if possible in cooperation with other donors, to support the development of a set of “meso-level” indicators that would be the best focus for USAID’s efforts to track and assess countries’ progress or problems with democratization. This would be a long-term and expensive effort, but there are already substantial numbers of candidate indicators that could potentially contribute to such an index (see, e.g., the review by Landman 2003). If the United States and other donors are going to continue to support the development of democracy worldwide, the committee strongly believes that it is time to invest the resources needed to provide high-quality indicators comparable to those that have been developed over time in other economic and social fields. Whether or not SORA or the Office of Democracy and Governance became the home for such an effort, its recent experience with a major quantitative assessment of the impact of U.S. democracy assistance (Finkel et al 2007, 2008) and its understanding of the needs of DG officers in Washington and in the field would make it a logical place from which such an initiative could be developed.

Improving Monitoring and Evaluation

This chapter has outlined the proposed evaluation initiative the committee believes should be the core of the effort to improve USAID’s ability to assess the effectiveness of its projects in the future. The committee’s recommendations for high-level leadership would support day-to-day implementation of the initiative and provide a central focus. One of the frequent comments that the field teams heard from DG officers was the desire for advice and assistance in understanding and developing impact evaluations, and this is a role SORA could readily play. It would also be a logical starting point if the recommendations for a wider effort to restore USAID’s evaluation capacity were implemented (NRC 2006:90-91). SORA could also be given responsibility for developing the social sciences advisory group and the broader partnerships with universities

that the committee recommends. These could both contribute to the work of the evaluation initiative and support learning from retrospective case studies.

Active Learning

While it will take time for the results of the evaluation initiative to mount and provide evidence for the positive or negative impact of various USAID DG projects, USAID can and should take advantage of other avenues to learn about DG assistance. The case studies and other analyses recommended in this report would be an essential part of this effort, as would regular opportunities to discuss DG officers' experiences and academic research on democratization. Active organizational learning means much more than simply having such research materials available for DG staff to peruse or view on the Web. As discussed in Chapter 8, it means having DG staff actively engaged with such materials through discussions and meetings with the authors of such research, probing to seek the lessons contained in the research. The continuing pilot effort for the "Voices from the Field" project discussed in Chapter 8 could over time become a key instrument in acquiring and disseminating insights from active practitioners as another element in this commitment to learning.

The committee thus recommends that part of the agenda for the Office of Democracy and Governance and the final part of the DG evaluation initiative should be a provision for active learning through regular meetings of DG staff with academics, NGOs, and think tank researchers who are exploring such issues as trajectories of democracy, the progress of democracy in various regions or nations, and the reception of DG programs in various settings. These need not all be in Washington but could include meetings in the field focused on regional issues or certain types of DG programs (e.g., having a conference in Africa on anticorruption programs that draws in regional DG staff). The planning for such meetings could involve partnerships with academics, think tanks, local partners, or other DG assistance donors.

Taken together and supported by the leadership of USAID, the SORA program and the wider efforts of the DG office and USAID that are more broadly discussed throughout this report would provide USAID with the capacity to effectively evaluate and continuously improve its work to support democratic development.

ROLE OF CONGRESS AND THE EXECUTIVE BRANCH

USAID cannot undertake the evaluation initiative and other efforts recommended here alone. A significant barrier to change is the agency's

uneasy relationship with Congress and uncertainty regarding its evolving relationship with the other parts of the Executive Branch.

Across the world, and across the U.S. government, there are efforts to improve results, accountability, and organizational knowledge of foreign assistance. The committee hopes that the efforts of SORA and the recommendations in this report will form part of this broad movement to reform foreign aid.

However, such improvement will only come with a commitment to learning what works and what does not, in a spirit that avoids blame and offers credit for learning that advances the effectiveness of aid. Military and medical institutions have learned that simply punishing failures leads to efforts to hide or cover up problems and thus to those problems being prolonged. Greater progress toward the overall goals is obtained when people are encouraged to report unintended problems or setbacks and are not penalized for them. Congress and the Executive Branch must take a position on foreign aid that learning of a program's ineffectiveness, although it may lead to ending that particular program, will not be used to undermine foreign aid in general or those who worked on that program. Indeed, given the currently uncertain knowledge and difficult challenge of advancing democracy in diverse conditions, learning that half or two-thirds of USAID's DG programs have real and significant effects in helping countries advance should be seen as fundamentally positive and evidence of success, while learning *which half* or one-third of programs are not effective should be seen as an important step in advancing the targeting and effectiveness of democracy assistance. Unrealistic expectations for universal success or rapid advances, given USAID's modest budgets for DG assistance and the complexities and many countervailing forces that prevail in the real world of democracy assistance, will not help the necessary learning—which will involve some incremental advances and some cases of learning from setbacks—that would lead to meaningful advances in the field of foreign assistance.

Congress, of course, is ultimately responsible for seeing that the public's money is used wisely, and it should be helped to understand that rigorous impact evaluations are an important tool in seeking that end. But more than that, the committee hopes for a renewed partnership between USAID and other branches of the federal government. Congress and Executive Branch policymakers should recognize that USAID DG programs cannot be held responsible for the successes or failures of democratic development in any given country. Even U.S. foreign policy as a whole with all of its instruments, of which USAID DG assistance is only a small part, may be unable to have a substantial impact. In turn, USAID should be held accountable for determining the success or failure of the DG projects it undertakes and for making a systematic effort to document and learn from what works and what does not. USAID should not fear

this process; repeated studies have now shown that, overall, democracy assistance *is effective* (Finkel et al 2007, 2008). What needs to be done next to improve such assistance is to learn more about which specific projects are being most effective and in what contexts. This simply cannot be done accurately without a strong commitment in *both* Congress and USAID to making sound impact evaluations a significant part of the agency's overall M&E and learning activities.

CONCLUSIONS

The committee wants to restate clearly its position that impact evaluations, especially randomized evaluations, though the most potent method of evaluating the true effects of DG projects where feasible and appropriate, are *not* the only important form of evaluation or the only path to improved DG programming. Process evaluations, debriefings, and sharing of personal insights among DG staff (e.g., "Voices from the Field"), as well as historical studies of democratic trajectories, also are essential components of knowledge building and improving DG activities. Yet perhaps the single most significant deficiency that the committee observed in regard to USAID learning which of its DG projects are most effective and when was the lack of well-designed impact evaluations of such projects. The committee sees an enormous opportunity for USAID to accelerate its learning and the effectiveness of its programming by learning through the proposed evaluation initiative whether and how impact evaluations could be applied to DG projects. More broadly, leadership that creates a strong expectation that high-quality evaluations are critical to USAID's future missions could improve USAID's global leadership in gaining knowledge about democracy promotion, give heightened credibility to USAID's relations with Congress, and—the committee believes—contribute greatly to achieving USAID's goals of supporting the spread and strengthening of democratic polities throughout the world.

REFERENCES

- Al-Momani, M.H. 2003. Financial Transfer and Its Impact on the Level of Democracy: A Pooled Cross-Sectional Time Series Model. Unpublished Ph.D. thesis, University of North Texas.
- Azpuru, D., Finkel, S., Pérez-Liñán, A., and Seligson, M.A. 2008. Trends in Democracy Assistance: What Has the United States Been Doing? *Journal of Democracy* 91(2):150-159.
- Blair, H., and Hanson, G. 1994. Weighing in on the Scales of Justice: Strategic Approaches for Donor-Supported Rule of Law Programs. USAID Program and Operations Assessment Report No. 7. Washington, DC: USAID Center for Development Information and Evaluation. Available at: http://www.usaid.gov/our_work/democracy_and_governance/publications/pdfs/pnaax280.pdf. Accessed on August 18, 2007.

- Bollen, K., Paxton, P., and Morishima, R. 2005. Assessing International Evaluations: An Example from USAID's Democracy and Governance Programs. *American Journal of Evaluation* 26:189-203.
- Epstein, S., Serafino, N., and Miko, F. 2007. *Democracy Promotion: Cornerstone of U.S. Foreign Policy?* Washington, DC: Congressional Research Service.
- Finkel, S.E., Pérez-Liñán, A., and Seligson, M.A. 2007. The Effects of U.S. Foreign Assistance on Democracy Building, 1990-2003. *World Politics* 59(3):404-439.
- Finkel, S.E., Pérez-Liñán, A., Seligson, M.A., and Tate, C.N. 2008. Deepening Our Understanding of the Effects of U.S. Foreign Assistance on Democracy Building: Final Report. Available at: <http://www.LapopSurveys.org>.
- HELP Commission. 2007. Beyond Assistance: The HELP Commission Report on Foreign Assistance Reform. Available at: <http://helpcommission.gov/>. Accessed on February 23, 2008.
- Hyman, G. 2008. Assessing Secretary of State Rice's Reform of U.S. Foreign Assistance. Carnegie Papers. Washington, DC: Carnegie Endowment for International Peace.
- Kalyvitis, S.C., and Vlachaki, I. 2007. Democracy Assistance and the Democratization of Recipients. Available at: <http://ssrn.com/abstract=888262>.
- Lancaster, C. 2000. *Transforming Foreign Aid: United States Assistance in the 21st Century*. Washington, DC: Peterson Institute for International Economics.
- Lancaster, C. 2006. *Foreign Aid: Diplomacy, Development, Domestic Policies*. Chicago: University of Chicago Press.
- Landman, T. 2003. Map-Making and Analysis of the Main International Initiatives on Developing Indicators on Democracy and Good Governance. Final Report. University of Essex. Available at: <http://www.oecd.org/dataoecd/0/28/20755719.pdf>. Accessed on April 27, 2008.
- National Endowment for Democracy. 2006. *The Backlash Against Democracy Assistance*. Washington, DC: National Endowment for Democracy.
- NRC (National Research Council). 2006. *The Fundamental Role of Science and Technology in International Development: An Imperative for the U.S. Agency for International Development*. Washington, DC: The National Academies Press.
- de Zeeuw, J., and Kumar, K. 2006. *Promoting Democracy in Postconflict Societies*. Boulder: Lynne Rienner.

Glossary

Terms in italics are defined elsewhere in the Glossary.

METHODOLOGICAL TERMS

Case: A spatially delimited phenomenon observed at a single point in time or over some period of time—for example, a political or social group, institution, or event. By construction, a case lies at the same level of analysis as the principal inference. Thus, if an inference pertains to the behavior of nation-states, cases in that study will be comprised of nation-states. An individual case may also be broken down into one or more *observations*, sometimes referred to as *within-case* observations.

Case study: The intensive study of a single case for the purpose of understanding a larger class of similar units (a *population* of cases). Note that while “case study” is singular—focusing on a single unit—a “case study research design” may refer to a study that includes several cases (e.g., comparative-historical analysis or the comparative method). Synonym: within-case analysis.

Causal inference: Determining from data whether—minimally—a causal factor (X) is thought to raise the probability of an effect (Y) occurring.

Control: See *Experiment*.

Experiment: Generically, a research design in which the causal factor of interest (the treatment or *intervention*) is manipulated by the researcher so

as to produce a more tractable analysis. Within social sciences circles the term is often equated with a research design in which an additional attribute obtains: Cases are randomized across treatment and control groups. Antonym: observational.

External validity: See *Validity*.

Internal validity: See *Validity*.

N: See *Observation*.

Observation: The most basic element of any empirical endeavor. Any piece of evidence enlisted to support a proposition. Conventionally, the number of observations in an analysis is referred to by the letter *N*. Confusingly, *N* is also used to refer to the number of *cases*.

Randomization: A process by which cases in a sample are chosen randomly (with respect to some subject of interest). An essential element for experiments that use control groups since the treatment and control groups, prior to treatment, must be similar in all respects that are relevant to the inference, and the easiest way to achieve this is through random selection. Sometimes, randomization occurs across matched pairs or within substrata of the sample (stratified random sampling), rather than across the entire population.

Research design: The way in which empirical evidence is brought to bear on a hypothesis.

Treatment: See *Experiment*.

Validity: Internal validity refers to the correctness of a hypothesis with respect to the sample (the cases actually studied by the researcher). External validity refers to the correctness of a hypothesis with respect to the population of an inference (cases not studied but that the inference is thought to explain). The key element of external validity thus rests on the representativeness of a sample—that is, its relative bias.

Variable: An attribute of an observation or a set of observations. In the analysis of causal relations, variables are understood either as independent (explanatory or exogenous), denoted *X*, or as dependent (endogenous), denoted *Y*.

Within-case analysis: See *Case study*.

X: See *Variable*.

Y: See *Variable*.

TYPES OF INTERVENTIONS

NOTE: USAID does not have a standard terminology to describe the various levels of activities it undertakes.

Activity: An intervention of a single type (e.g., training judges).

Intervention: Any activity or set of activities (e.g., project, program) undertaken by a funder. Usually employed in the context of an *evaluation*; here, the intervention is the independent variable whose effect on a policy outcome is being assessed.

Program: Includes all projects that address a particular USAID policy area, such as democracy and governance, health, or humanitarian assistance.

Project: Includes all activities within the scope of a particular contract or grant.

TYPES OF APPRAISALS

Country assessment: Appraisal of policy performance at the country level (e.g., levels of corruption or quality of democracy). Purposes of country assessments include tracking progress and regress across countries (including democratic and authoritarian transitions), identifying common patterns of transition and, possibly, the causal drivers of transition. This information should help funders decide in which countries investments might be most productive and also the sectors of a country that are most in need of assistance. Measured by *meso-* and *macro-level indicators*.

Evaluation: See below.

Monitoring: Routine oversight of a project's implementation (e.g., whether funds are spent properly and other terms of the contract are adhered to). Usually measured with *outputs* (e.g., number of judges trained).

Strategic: Appraisal of the opportunities and constraints in various countries for transition to democracy or the stabilization or better functioning of democracy. Should be based on hypotheses about the factors that drive or inhibit democracy in specific contexts. Strategic appraisals guide USAID's central decisions on how much democracy assistance to allot to specific countries in specific time periods. *Country assessments*, made by USAID DG missions, also involve a strategic appraisal.

Tactical: Appraisal of which programs should be employed, in which areas or sectors, to best assist a country's transition to, or stabilization of, democracy. Tactical decisions are generally made at the level of the USAID mission DG office, following a *country assessment*. Good tactical decisions

depend on accumulated knowledge about the impacts of specific DG programs in particular contexts, gained through good *evaluations*.

TYPES OF EVALUATIONS

NOTE: Evaluations should be considered one type of appraisal.

Impact evaluation: A study of a project or set of projects that seeks to determine how observed outcomes differ from what most likely would have happened in the absence of the project(s) by using comparison or control groups or random assignment of assistance across groups or individuals to provide a reference against which to assess the observed outcomes for groups or individuals who received assistance. Randomized designs offer the most accuracy and credibility in determining program impacts and therefore should be the first choice, where feasible, for impact evaluation designs. However, such designs are not always feasible or appropriate, and a number of other designs also provide useful information, but with diminishing degrees of confidence, for determining the impact of many different kinds of assistance projects.

Output evaluation (generally equivalent to “project monitoring” within USAID): These evaluations consist of efforts to document the degree to which a program has achieved certain targets in its activities. Targets may include spending specific sums on various activities, giving financial support or training to a certain number of nongovernmental organizations (NGOs) or media outlets, training a certain number of judges or legislators, or carrying out activities involving a certain number of villagers or citizens. Output evaluations or monitoring are important for ensuring that activities are carried out as planned and that money is spent for the intended purposes.

Participatory evaluation: Individuals, groups, or communities that will receive assistance are involved in the development of project goals, and investigators interview or survey participants during and/or after a project was carried out to determine what their goals and expectations are for the project, how valuable the activity was to them, and whether they were satisfied with the project’s results.

Process evaluation: Focuses on how and why a program unfolded in a particular fashion, and if there were problems, on why things did not go as originally planned. Usually conducted after completion of a project, often using teams of experts who conduct interviews and examine project records. Currently the primary source of “lessons learned” and “best practices” intended to inform and assist project managers and implementers.

TYPES OF INDICATORS

Indicator: Generically, any operational measure of an underlying concept. May be measured at local, regional, or national levels. Usually quantitative in nature, although may be formed from data originally gathered in a qualitative format. Includes *outputs*, *outcomes*, and *meso-* and *macro-level indicators*, as discussed below. For USAID's purposes, good indicators are valid (the measurement is in accordance with the underlying concept), cross-nationally comparable, and reliable (different applications of the indicator result in similar if not identical measurements).

Macro-level indicators: Measure country-level features at a highly aggregated level (e.g., democracy). Used for *country assessment*.

Meso- or sectoral-level indicators: Measure country-level features in some rather specific policy area (e.g., elections). Used for *country assessment* and, very occasionally, for program/project *evaluation*.

Outcomes: Measure the impact of an intervention on some aspect of society. Used for program/project *evaluation*.

Outputs: Measure the specific targets of a program. Often used for program/project *monitoring*

SUBSTANTIVE CONCEPTS

Authoritarian regimes: Governments in which leaders are not chosen by competitive elections and in which all political opposition is repressed. All media, local government, judiciary, and legislature are tightly controlled by the executive.

Democracy: Generally, rule by the people; also known as popular sovereignty; an aspect of *governance*. In reaching for a more specific definition, two general strategies may be identified. Minimalist definitions usually center on the idea of contestation (competition). Maximalist (ideal-type) definitions add additional qualifiers such as liberty/freedom, accountability, responsiveness, deliberation, participation, political equality, and social equality.

Full democracy: A system of government in which leaders are chosen by open and fair electoral competition and in which all of the political liberties and rights needed to ensure such open and fair competition—personal security and nondiscrimination, rule of law, accountability of officials, civilian control, and freedom of speech, assembly, and media—are well institutionalized and protected.

Governance: The quality of government (e.g., rule of law, low corruption, high efficiency, high performance on dimensions deemed valuable for improving human welfare). May include some or all features of *democracy*.

Partial democracy: A system of government in which leaders are chosen by electoral competition, but such competition is not fully open or fair, and in which many of the political liberties and rights needed to ensure open and fair competition are absent or irregular. Elections are often marked by violence or disorders, elected officials are not fully accountable, and certain groups may be excluded from politics or disadvantaged by state control of media or electoral procedures.

Semiauthoritarian regimes: Governments in which leaders are not chosen by competitive elections but in which some political liberties are allowed. Leaders do stand for elections, but the eligibility and activities of the opposition are so tightly constrained that the outcome is never in doubt. There may be some independent media, some opposition political parties, and some diversity of representation in parliament or local governments. There may be some elements of the judiciary or electoral monitoring that function with autonomy.

Appendixes

A

Biographical Sketches of Committee Members

Jack A. Goldstone—(Chair), George Mason University

Jack A. Goldstone is the Virginia E. and John T. Hazel Jr. Professor at the George Mason School of Public Policy and a senior research scholar at the Mercatus Institute. His work on social movements, revolutions, democratization, and economic growth has won the Distinguished Scholarly Publication Award of the American Sociological Association, and Fellowships from the U.S. Institute for Peace, the MacArthur Foundation, and the Center for Advanced Study in the Behavioral Sciences. He is a senior member of the Political Instability Task Force and is director of the Center for Global Policy at George Mason. The author or editor of nine books and nearly 100 research articles, Goldstone is a consultant to the U.S. State Department, intelligence agencies, and the United States Agency for International Development (USAID). His areas of expertise include revolutions and social movements, comparative economic development, comparative politics, conflict and rebellion, democracy, fragile states, and political demography.

Larry Garber, New Israel Fund

Larry Garber joined the New Israel Fund following five years as director of USAID's West Bank and Gaza mission. Previously, he was senior policy advisor and deputy assistant administrator of the Bureau of Policy and Program Coordination at USAID. Garber was a senior associate at the National Democratic Institute for International Affairs from 1988 to 1993, organizing international election observer missions leading to historic

governmental transitions in the Philippines, Chile, Pakistan, Panama, Bulgaria, and Zambia. He has also served as an advisor to a number of governments on election law reform issues. Garber served as legal director of the International Human Rights Law Group from 1983 to 1988, preparing the first-ever guide for international election observers. He has served as a member of the adjunct faculty of the Washington College of Law of American University and as a consultant to the United Nations, the Organization of American States, and the Organization for Security and Cooperation in Europe. Garber is a 1980 graduate of Columbia University with a joint J.D. and M.A. in international affairs. He received his B.A. in 1976, from Queens College of the City University of New York and spent a year of his undergraduate studies at the Hebrew University in Jerusalem.

John Gerring, Boston University

John Gerring received his Ph.D. from the University of California at Berkeley in 1993. He is currently an associate professor of political science at Boston University where he teaches courses on methodology and comparative politics. His books include *Party Ideologies in America, 1828-1996* (Cambridge University Press, 1998), *Social Science Methodology: A Criterial Framework* (Cambridge University Press, 2001), *Case Study Research: Principles and Practices* (Cambridge University Press, 2007), *Global Justice: A Prioritarian Manifesto* (under review), and *Centripetalism: A Theory of Democratic Governance* (with Strom Thacker; under review), *Concepts and Method: Giovanni Sartori and His Legacy* (with David Collier; under review), and *Democracy and Development: A Historical Perspective* (with Strom Thacker; in process).

His articles have appeared in the *American Political Science Review*, *British Journal of Political Science*, *International Organization*, *Journal of Policy History*, *Journal of Theoretical Politics*, *Party Politics*, *Political Research Quarterly*, *Polity*, *Social Science History*, *Studies in American Political Development*, and *World Politics*. He was a fellow of the School of Social Science at the Institute of Advanced Study (2002-2003). He is the former editor of *Qualitative Methods: Newsletter of the American Political Science Association Organized Section on Qualitative Methods* and president-elect of the Qualitative Methods section.

Clark C. Gibson, University of California, San Diego

Clark Gibson is professor of political science and director of the International Studies Program at University of California, San Diego. He studies the politics of development, democracy, and the environment. He has explored issues related to these topics in Africa, Central and South America, and the United States. The results of his work have appeared

in journals such as *Comparative Politics*, *World Development*, *Annual Review of Political Science*, *Social Science Quarterly*, *Human Ecology*, *Conservation Biology*, *Ecological Economics*, and *African Affairs*. Gibson's research about the politics of wildlife policy in Africa appears in his book, *Politicians and Poachers: The Political Economy of Wildlife Policy in Africa*. He has also co-edited two volumes: *People and Forests: Communities, Institutions, and Governance*, which uses techniques from the natural and social sciences to examine the local governance of forests; and *Communities and the Environment: Ethnicity, Gender, and the State in Community-Based Conservation*, which explores the complex and multilayered links between members and their natural resources. Gibson's latest coauthored book, *Samaritan's Dilemma: The Political Economy of Development Aid*, analyzes the political economy of foreign aid and offers suggestions for its improvement. His current research focuses on the accountability between governments and citizens in Africa.

Mitchell A. Seligson, Vanderbilt University

Mitchell A. Seligson is the Centennial Professor of Political Science at Vanderbilt University and is also a fellow at the Center for the Americas at Vanderbilt. He founded and directs the Latin American Public Opinion Project (LAPOP). LAPOP has conducted over 60 surveys of public opinion, mainly focused on democracy, in many countries in Latin America, but more recently has included projects in Africa and the Balkans. Prior to joining the faculty at Vanderbilt, he held the Daniel H. Wallace Chair of Political Science at the University of Pittsburgh, and also served there as director of the Center for Latin American Studies. He has held grants and fellowships from the Rockefeller Foundation, the Ford Foundation, the National Science Foundation, the Howard Heinz Foundation, Fulbright, USAID, and others, and has published over 80 articles and more than a dozen books and monographs. In addition to consulting for USAID, he also consults for the World Bank, the United Nations Development Program, and the Inter-American Development Bank. His most recent books are *Elections and Democracy in Central America, Revisited*, co-edited with John Booth, and *Development and Underdevelopment, the Political Economy of Global Inequality* (3rd ed., 2003), co-edited with John Passé-Smith.

Jeremy M. Weinstein, Stanford University

Jeremy M. Weinstein is Assistant Professor of Political Science at Stanford University, an affiliated faculty member at the Center for Democracy, Development, and the Rule of Law and the Center for International Security and Cooperation, and a nonresident fellow at the Center for Global Development. His research focuses on civil wars and communal violence, ethnic politics and the provision of public goods, postconflict reconstruc-

tion, and democracy promotion. He is the author of *Inside Rebellion: The Politics of Insurgent Violence* (Cambridge University Press, 2007). He has also published articles in the *American Political Science Review*, *Journal of Conflict Resolution*, *Foreign Affairs*, *Journal of Democracy*, *World Policy Journal*, and the *SAIS Review*. Previously, Weinstein directed the bipartisan Commission on Weak States and U.S. National Security. He has also worked on the National Security Council staff, served as a visiting scholar at the World Bank, and held fellowships at the Woodrow Wilson International Center for Scholars and the Brookings Institution. He is a term member of the Council on Foreign Relations. Weinstein received a B.A. with high honors from Swarthmore College, and an M.A. and Ph.D. in political economy and government from Harvard University.

B

Committee Meetings and Participants

Committee on the Evaluation of USAID Democracy Assistance Programs

August 29-30, 2006

AGENDA

August 29th

CLOSED SESSION

8:30–10:15

OPEN SESSION

10:15 Meeting begins

- Opening remarks by committee chair
- Introduction of committee members
- Brief project overview
- Plan for the meeting

10:30 General Remarks

Overview of USAID's DG programs

- USAID's DG strategy

- Typical programs
- Coordination with other U.S. government agencies and donors

Open discussion with committee members

Overview of USAID Evaluation and Performance Management

- Budget and strategic planning
- Assessments and evaluations
- Performance management plans

12:00 Lunch

1:00 Overview of the Strategic and Operational Research Agenda (SORA)

- Introduction and background on SORA
- Previous studies, including the Quantitative Study
- Expectations for this study

Open discussion with committee members

1:45 Discussion with USAID of Key Committee Tasks

- Research design
- Pilot field studies
- Future research

3:00 Break

CLOSED SESSION

3:15–5:00

August 30th

OPEN SESSION

8:30 Continental breakfast available in the meeting room

9:00 Meeting begins

- Plans for the day
- Plans for Field Study #1

10:30 Break

CLOSED SESSION

10:45–4:00

PARTICIPANTS

Patricia Alexander
U.S. Agency for International
Development

Tabitha Benney
National Academies

Mark Billera
U.S. Agency for International
Development

Richard Bissell
National Academies

David Black
U.S. Agency for International
Development

Paul Bonicelli
U.S. Agency for International
Development

John Boright
National Academies

Ed Connerley
U.S. Agency for International
Development

Jeffrey Fillar
Booz Allen Hamilton

Pat Fn'Piere
U.S. Agency for International
Development

Larry Garber
New Israel Fund

John Gerring
Boston University

Clark C. Gibson
University of California, San
Diego

Jack A. Goldstone, *Chair*
George Mason University

Andrew Green
U.S. Agency for International
Development

Rita Guenther
National Academies

April Hahn
U.S. Agency for International
Development

Gary Hansen
U.S. Agency for International
Development

Jo Husbands
National Academies

Jerry Hyman
U.S. Agency for International
Development

Josh Kaufman
U.S. Agency for International
Development

Eric Kite
U.S. Agency for International
Development

Neil Levine
U.S. Agency for International
Development

Kimberly Ludwig
U.S. Agency for International
Development

Cathy Niarchos
U.S. Agency for International
Development

Maria Rendon-Labadan
U.S. Agency for International
Development

Margaret Sarles
U.S. Agency for International
Development

Keith Schulz
U.S. Agency for International
Development

Mitchell A. Seligson
Vanderbilt University

Paul Stern
National Academies

**Committee on the Evaluation of USAID
Democracy Assistance Programs**

September 18-19, 2006

September 18th

CLOSED SESSION

September 19th

CLOSED SESSION

8:30–10:30

OPEN SESSION

10:30 Discussion with Democracy and Governance Practitioners

Participants:

Richard McCall, Senior Vice President for Programs,
Creative Associates

Kim Mahling Clark, Senior Associate, Creative Associates

Michael Lund, Management Systems International, Inc. and
Woodrow Wilson International Center for Scholars

Taras Kuzio, Senior Transatlantic Fellow, the German
Marshall Fund

Michelle Bekkering, International Republican Institute
(Invited)

Rakesh Sharma, IFES (Invited)

12:30 Working Lunch/Continued Discussion with Democracy and Governance Practitioners

Kenneth Wollack, President, National Democratic Institute

Christopher Fomunyoh, Senior Associate for Africa and Regional Director for Central and West Africa, NDI

3:30–3:45 Break

CLOSED SESSION

3:45–5:00

PARTICIPANTS

Tabitha Benney
National Academies

Mark Billera
U.S. Agency for International
Development

Richard Bissell
National Academies

David Black
U.S. Agency for International
Development

John Boright
National Academies

Kim Mahling Clark
Creative Associates

Christopher Fomunyoh
NDI

Larry Garber
New Israel Fund

John Gerring
Boston University

Clark C. Gibson
University of California, San
Diego

Jack A. Goldstone, *Chair*
George Mason University

Andrew Green
Georgetown University

Rita Guenther
National Academies

Jo Husbands
National Academies

Jerry Hyman
U.S. Agency for International
Development

Taras Kuzio
The German Marshall Fund

Michael Lund
Management Systems
International, Inc. and
Woodrow Wilson International
Center for Scholars

Richard McCall
Creative Associates

Margaret Sarles
U.S. Agency for International
Development

Ramziya Shakirova
George Mason University

Jeremy Weinstein
Stanford University

Rakesh Sharma
IFES

Susan Wolchik
George Washington University

Mitchell A. Seligson
Vanderbilt University

Kenneth Wollack
NDI

Paul Stern
National Academies

**Committee on the Evaluation of USAID
Democracy Assistance Programs**

November 9-10, 2006

Thursday, November 9

OPEN SESSION

- 8:30 Continental breakfast available in the meeting room
- 9:00 Meeting begins
- Opening remarks by committee chair
 - Plan for the meeting
- 9:15 Session #1: Rule of Law
- 10:15 Break
- 10:30 Session #2: Governance
- 11:30 Session #3: Civil Society
- 12:30 Working Lunch
- 1:15 Session #4: Elections and Processes
- 2:15 Break
- 2:30 General discussion
- 4:30 Summary
- 5:00 Adjourn
- 6:00 Committee Working Dinner

Friday, November 10th

CLOSED SESSION

PARTICIPANTS

Tabitha Benney National Academies	Michael Henning Department of State
Richard Bissell National Academies	Jo Husbands National Academies
Eric Bjornlund Democracy International	Jerry Hyman U.S. Agency for International Development
David Black U.S. Agency for International Development	Cathy Niarchos U.S. Agency for International Development
John Boright National Academies	Bhavani Pathak U.S. Agency for International Development
Sharon Carter U.S. Agency for International Development	Bea Reaud American University
Glenn Cowan Democracy International	Maria Rendon-Labadan U.S. Agency for International Development
Bill Gallery Democracy International	Margaret Sarles U.S. Agency for International Development
Larry Garber New Israel Fund	Mitchell A. Seligson Vanderbilt University
John Gerring Boston University	Paul Stern National Academies
Clark C. Gibson University of California, San Diego	Kathryn Stratos U.S. Agency for International Development
Jack A. Goldstone, <i>Chair</i> George Mason University	Jeremy Weinstein Stanford University
Andrew Green Georgetown University	Shawna Wilson U.S. Agency for International Development
Rita Guenther National Academies	
Sean Hall U.S. Agency for International Development	

**Committee on the Evaluation of USAID
Democracy Assistance Programs**

May 3-4, 2007

May 3rd

CLOSED SESSION

8:30–10:30

OPEN SESSION

10:45 Update from USAID on Field Visits

- Field visit schedules
- Mission issues and updates
- Continued discussion of the revised field plan

11:15 Public Workshop Issues

11:30 Voices from the Field

12:30 Working Lunch (NAS Cafeteria)

1:30 Review of Project Statement and Deliverables

1:45 NRC Report Review Process

2:30 Break

CLOSED SESSION

2:45–5:00 Meeting Adjourns

May 4th

CLOSED SESSION

8:30–5:00 Meeting Adjourns

PARTICIPANTS

Tabitha Benney
National Academies

Mark Billera
U.S. Agency for International
Development

Richard Bissell
National Academies

David Black
U.S. Agency for International
Development

John Boright
National Academies

Larry Garber
New Israel Fund

John Gerring
Boston University

Clark C. Gibson
University of California, San
Diego

Jack A. Goldstone, *Chair*
George Mason University

Andrew Green
World Justice Project
American Bar Association

Rita Guenther
National Academies

Jo Husbands
National Academies

Dan Posner (*by teleconference*)
University of California, Los
Angeles

Margaret Sarles
U.S. Agency for International
Development

Mitchell A. Seligson
Vanderbilt University

Paul Stern
National Academies

Jeremy Weinstein (*by teleconference*)
Stanford University

Committee on the Evaluation of USAID Democracy Assistance Programs

July 19-20

July 19th

OPEN SESSION

9:45 Reports from the Field Visits: A Conversation with USAID
and the Committee on Evaluation of USAID Democracy
Assistance Programs

Project consultants, and USAID implementers

- Brief reports from the field visits
- Initial ideas for improving project evaluation
- Challenges, obstacles, and opportunities

10:45 Break

11:00 Discussion continues

12:45 Public Session adjourns; Working lunch for committee in
cafeteria

CLOSED SESSION

1:30–5:00

July 20th**OPEN SESSION**

- 8:30 Continental breakfast available in the meeting room
- 9:00 Updates and conversation with USAID
- Issues and questions from yesterday's discussion

CLOSED SESSION

10:30–5:00

PARTICIPANTS

Moises Arce Consultant	Thad Dunning (<i>by teleconference</i>) Consultant
Aaron Azelton National Democratic Institute	Matt Dippell National Democratic Institute
Tabitha Benney National Academies	Patrick Elliot National Democratic Institute
Mark Billera U.S. Agency for International Development	Larry Garber New Israel Fund
Richard Bissell National Academies	John Gerring Boston University
David Black U.S. Agency for International Development	Clark C. Gibson University of California, San Diego
John Boright National Academies	Jack A. Goldstone, <i>Chair</i> George Mason University
Don Chisholm USAID Democracy Fellow, Rule of Law	Rita Guenther National Academies
International Judicial Relations Committee	Jo Husbands National Academies
Brionne Dawson National Democratic Institute	Gerald Hyman Center for Strategic and International Studies
Ed Dennison Development Associates	Lisa Klimas National Democratic Institute

Jerry Lavery
National Democratic Institute

Devra Moehler
Consultant

Geetha Nagarajan
IRIS

Dan Posner (*by teleconference*)
Consultant

Margaret Sarles (*by teleconference*)
U.S. Agency for International
Development

Stephen Schwenke
Creative Associates International
Inc.

Louis Siegel
ARD, Inc.

Cecelia Skott
SUNY Center for International
Development

Mitchell A. Seligson
Vanderbilt University

Paul Stern
National Academies

Jeremy Weinstein
Stanford University

C

Measuring Democracy

This appendix contains three sections to support and expand the material in Chapter 3. The statistical analysis presented in the first section was carried out by Ramziya Shakirova, a graduate student at George Mason University, on behalf of the committee. The second section contains the agenda and participants list for a committee workshop, “Democracy Indicators for Democracy Assistance,” held at Boston University in January 2007. The last section is an “Outline for a Disaggregated Meso-level Democracy Index” by John Gerring, which contains additional material related to the index proposed in Chapter 3.

STATISTICAL ANALYSIS

Spearman vs. Pearson Coefficients

The comparison of Spearman and Pearson correlation coefficients shows that on the whole, they are quite similar. However, in some cases the Spearman correlation coefficients are not significant (probably, the Pearson coefficients are too, but Stata does not display a significance level for the Pearson coefficients), which means that the Freedom House (FH) and Polity scores are in fact independent. The countries in “Partially Free Group” with insignificant correlations are:

Cambodia: Pearson is 0.3281; Spearman is 0.3453, not significant

Armenia: Pearson is 0.1632; Spearman is 0.1615, not significant

Azerbaijan: Pearson is -0.0808 ; Spearman is 0.2864 , but not significant
 Moldova: Pearson is 0.6019 ; Spearman is 0.4550 , not significant
 Ukraine: Pearson is -0.3344 ; Spearman is -0.3015 , not significant
 Afghanistan: Pearson is 0.1832 ; Spearman is 0.2388 , not significant
 Egypt: Pearson is -0.2036 ; Spearman is -0.0889 , but not significant
 Yemen: Pearson is -0.0096 ; Spearman is -0.2060 , not significant
 Tunisia: Pearson is -0.0265 ; Spearman is -0.0452 , not significant
 Mexico: Pearson is 0.4544 ; Spearman is 0.2681 , not significant
 Greece: Pearson is 0.896 ; Spearman is 0.1609 , not significant
 Macedonia: Pearson is 0.373 ; Spearman is 0.3924 , not significant
 Sierra Leone: Pearson is 0.5094 ; Spearman is 0.2858 , not significant
 Zimbabwe: Pearson is 0.2791 ; Spearman is 0.2612 , but not significant
 Burundi: Pearson is 0.4269 ; Spearman is 0.2823 , not significant
 Cameroon: Pearson is -0.1538 ; Spearman is -0.1018 , not significant
 Comoros: Pearson is -0.0408 ; Spearman is 0.2358 , not significant
 Kenya: Pearson is 0.1287 ; Spearman is -0.1646 , not significant

For two countries (Colombia and Côte d'Ivoire), the coefficients are close in their magnitude, although the Spearman coefficients are significant only at the 10 percent level, but not the 5 percent level.

Correlation of First Differences

The average correlation coefficients for the first differences in the group of **"Partially Free"** countries are low for the **Former Soviet Union** and the **Middle East** (Table C-1).

- For the Former Soviet Union, the average correlation coefficient is equal to 0.148 . Particularly, the coefficients are low for Armenia (0.1871) and Tajikistan (0.1320), and close to zero for the Ukraine (0.0891). Negative coefficients are observed for Kazakhstan (-0.1562), Moldova (-0.1800), and Russia (-0.5188). Satisfactory coefficients for the first differences are found in this group only for Azerbaijan, Belarus, and Georgia.
- For the Middle East, the average is 0.285392 . In this group, negative coefficients are observed for Egypt (-0.1292), Iran (-0.0531), and Yemen North (-0.0408), and close to zero, but positive, for Tunisia (0.0840).

In other regional groups there are also some countries with negative or zero coefficients: (Malaysia (-0.083), Panama (-0.2525), Angola (-0.0788), Côte d'Ivoire (-0.091), Liberia (0.000), Madagascar (0.0589), Rwanda (-0.2813), Togo (-0.0195), Uganda (0.0211), Chad (-0.218), Comoros (-0.3467), and Equatorial Guinea (-0.3725). The average correlation

coefficients for other regional groups are the following: Asia (0.4829), Latin America (0.54092105), and Africa (0.408083).

In the group of “**Democratic**” countries (Table C-2), negative correlations for the first differences are observed for Cyprus (−0.6930), France (−0.0197), and Mauritius (−0.0197), and close to zero coefficient for Trinidad (0.0163). The average for this group is also very low, and equal to 0.11855714.

For “**Autocratic**” countries (Table C-3), negative coefficients are observed for China (−0.0113), Oman (−0.0496), Yemen South (−0.4123), and Mauritania (−0.0197), and zero correlation for Syria.

There are several countries where the correlation coefficients for the first differences are positive, although correlations between FH and Polity scores are negative (Bahrain, Iraq, and Morocco).

The average correlation coefficient for “autocratic” countries is 0.296829.

TABLE C-1 “Partially Free” Countries (Polity Scores −5 to +7)—
Correlations with FH Scores

Country	Number of Observations	Years	Pearson Correlation Coefficient	Spearman Rank Order Coefficient	Correlation Coefficient for First Differences
Asia					
Cambodia	21	1972-1978; 1988-2002	0.32810	0.3453* ¹	0.2763
Fiji	31	1972-2002	0.86190	0.9474	0.6204
Indonesia	31	1972-2002	0.78120	0.6197	0.5473
Malaysia	31	1972-2002	0.64410	0.6942	−0.0830
Mongolia	31	1972-2002	0.98480	0.9694	0.6717
Philippines	31	1972-2002	0.93820	0.8965	0.5049
South Korea	31	1972-2002	0.96250	0.8702	0.4825
Taiwan	31	1972-2002	0.93100	0.9355	0.5105
Thailand	31	1972-2002	0.72970	0.6443	0.8155
Average			0.79572	0.769167	0.4829
Variance			0.04410	0.043645	0.06668676
Standard deviation			0.20999	0.208915	0.2582378
Former Soviet Union					
Armenia	11	1992-2002	0.16320	0.1615*	0.1871
Azerbaijan	11	1992-2002	−0.08080	0.2864*	0.6191
Belarus	11	1992-2002	0.97200	0.9877	0.7319
Georgia	11	1992-2002	0.81110	0.7709	0.4286
Kazakhstan	11	1992-2002	0.54070	0.6992	−0.1562
Moldova	11	1992-2002	0.60190	0.4550*	−0.1800

TABLE C-1 Continued

Country	Number of Observations	Years	Pearson Correlation Coefficient	Spearman Rank Order Coefficient	Correlation Coefficient for First Differences
Russia	11	1992-2002	-0.77130	-0.7287	-0.5188
Tajikistan	11	1992-2002	0.75710	0.7944	0.1320
Ukraine	11	1992-2002	-0.33440	-0.3015*	0.0891
Average			0.29550	0.347211	0.148
Variance			0.34806	0.317729	0.16145
Standard deviation			0.58997	0.563674	0.40181
Middle East					
Afghanistan	24	1972-2002 (7 missing values)	0.18320	0.2388*	0.3839
Algeria	31	1972-2002	0.59750	0.5458	0.6766
Bangladesh	31	1972-2002	0.83010	0.7977	0.5921
Egypt	31	1972-2002	-0.20360	-0.0889*	-0.1292
Iran	31	1972-2002	-0.43260	-0.4784	-0.0531
Jordan	31	1972-2002	0.87020	0.9045	0.3692
Yemen North	18	1972-1989	0.64240	0.5127	-0.0408
Yemen	13	1990-2002	-0.00960	-0.2060*	0.4714
Nepal	31	1972-2002	0.77640	0.7313	0.2268
Pakistan	31	1972-2002	0.81720	0.8421	0.6368
Sri Lanka	31	1972-2002	0.75830	0.7932	0.2070
Tunisia	31	1972-2002	-0.02650	-0.0452*	0.0840
Average			0.40025	0.378967	0.285392
Variance			0.21838	0.227563	0.07862994
Standard deviation			0.46731	0.477035	0.2804103
Latin America					
Argentina	31	1972-2002	0.84850	0.6961	0.5757
Bolivia	31	1972-2002	0.86500	0.7738	0.1415
Brazil	31	1972-2002	0.72550	0.6021	0.3323
Chile	31	1972-2002	0.97890	0.8875	0.8631
Colombia	31	1972-2002	0.38110	0.3248**	0.3981
Dominican Republic	31	1972-2002	0.41910	0.3863	0.5327
Ecuador	31	1972-2002	0.96720	0.8145	0.8878
Honduras	31	1972-2002	0.91660	0.5534	0.3793
Uruguay	31	1972-2002	0.96570	0.9126	0.8292
Venezuela	31	1972-2002	0.91890	0.8858	0.6999
Guatemala	31	1972-2002	0.52950	0.4147	0.7231
Guyana	31	1972-2002	0.61740	0.6223	0.4895
Haiti	31	1972-2002	0.77960	0.5672	0.8299
El Salvador	31	1972-2002	0.44230	0.4156	0.4314
Mexico	31	1972-2002	0.45440	0.2681*	0.1765

TABLE C-1 Continued

Country	Number of Observations	Years	Pearson Correlation Coefficient	Spearman Rank Order Coefficient	Correlation Coefficient for First Differences
Nicaragua	31	1972-2002	0.69900	0.7018	0.7508
Panama	31	1972-2002	0.78820	0.8845	-0.2525
Paraguay	31	1972-2002	0.93820	0.8731	0.8007
Peru	31	1972-2002	0.90660	0.8867	0.6885
Average			0.74430	0.656363	0.5409211
Variance			0.04356	0.046605	0.08946278
Standard deviation			0.20870	0.215881	0.2991033
European Union					
Turkey	31	1972-2002	0.45220	0.6856	0.5807
Spain	31	1972-2002	0.97670	0.8546	0.5638
Greece	31	1972-2002	0.89570	0.1609*	0.6981
Macedonia	11	1992-2002	0.37300	0.3924*	0.6713
Portugal	31	1972-2002	0.95720	0.8370	0.7478
Albania	31	1972-2002	0.95260	0.9623	0.2263
Bulgaria	31	1972-2002	0.99360	0.9750	0.9330
Croatia	12	1991-2002	0.92370	0.7648	0.5851
Czechoslovakia	21	1972-1992	0.99360	0.8176	0.9911
Yugoslavia	31	1972-2002	0.87510	0.7940	0.3907
Hungary	31	1972-2002	0.98420	0.9123	0.7045
Poland	31	1972-2002	0.98750	0.9314	0.6476
Romania	31	1972-2002	0.93910	0.8766	0.5623
Slovakia	10	1993-2002	0.90800	0.9039	0.5754
Average			0.87229	0.776314	0.6341214
Variance			0.03957	0.05302	0.03728278
Standard deviation			0.19893	0.230261	0.19308749
Africa					
Angola	26	1976-1991; 1993-2002	0.64430	0.6916	-0.0788
Côte d'Ivoire	31	1972-2002	0.21770	0.3316**	-0.0910
Kenya	31	1972-2002	0.12870	-0.1646*	0.5633
Liberia	25	1972-1989; 1996-2002	-0.05210	-0.0088	0.0000
Lesotho	30	1972-1997; 1999-2002	0.79260	0.8084	0.6109
Madagascar	31	1972-2002	0.90170	0.8745	0.0589
Malawi	31	1972-2002	0.99040	0.9925	0.9026
Mali	31	1972-2002	0.98390	0.8861	0.7004
Mozambique	28	1975-2002	0.95730	0.9106	0.6709
Nigeria	31	1972-2002	0.86180	0.7510	0.8536
Niger	31	1972-2002	0.94460	0.8791	0.7003

TABLE C-1 Continued

Country	Number of Observations	Years	Pearson Correlation Coefficient	Spearman Rank Order Coefficient	Correlation Coefficient for First Differences
Rwanda	31	1972-2002	-0.54360	-0.7643	-0.2813
South Africa	31	1972-2002	0.94290	0.8831	0.3889
Senegal	31	1972-2002	0.72580	0.6668	0.2634
Sierra Leone	27	1972-1996; 1998-2000	0.50940	0.2858*	0.7987
Sudan	31	1972-2002	0.54780	0.4058	0.8398
Tanzania	31	1972-2002	0.95040	0.8904	0.5234
Togo	31	1972-2002	0.81130	0.8327	-0.0195
Uganda	29	1972-1978; 1980-1984; 1986-2002	0.57990	0.6659	0.0211
Zambia	31	1972-2002	0.87650	0.8812	0.8956
Zimbabwe	31	1972-2002	0.27910	0.2612*	0.2902
Benin	31	1972-2002	0.98980	0.8889	0.9128
Burkina Faso	31	1972-2002	0.82510	0.8709	0.7828
Burundi	28	1972-1992; 1996-2002	0.42690	0.2823*	0.2181
Cameroon	31	1972-2002	-0.15380	-0.1018*	0.2964
Central African Republic	31	1972-2002	0.90760	0.8678	0.6987
Chad	26	1972-1977; 1984-2002	0.86520	0.8529	-0.218
Comoros	24	1976-1994; 1996-2002	-0.04080	0.2358*	-0.3467
Congo Brazzaville	31	1972-2002	0.88020	0.8362	0.7504
Equatorial Guinea	31	1972-2002	-0.43130	-0.4482	-0.3725
Ethiopia	29	1972-1973; 1974-1990; 1992-2002	0.83050	0.6848	0.1520
Gabon	31	1972-2002	0.92670	0.9295	0.6838
Gambia	31	1972-2002	0.92960	0.9447	0.8985
Ghana	31	1972-2002	0.91910	0.8917	0.5412
Guinea Bissau	27	1975-1997; 1999-2002	0.94260	0.8784	0.8501
Guinea	31	1972-2002	0.87330	0.9561	0.2320
Average			0.631697	0.598072	0.408083
Variance			0.182355	0.19153	0.165665
Standard deviation			0.427031	0.437642	0.40702

*Coefficient is not significant at 5 percent significance level.

**Coefficient is not significant at 5 percent level, but is significant at 10 percent level.

TABLE C-2 "Democratic" Countries (Polity Scores 8-10)—
Correlations with FH Scores

Country	Number of Observations	Years	Correlation Coefficient	Correlation for First Differences
India	31	1972-2002	0.17940	0.5833
Israel	31	1972-2002	0.34690	0.5708
Jamaica	31	1972-2002	0.20880	0.3919
Trinidad	31	1972-2002	0.15910	0.0163
Cyprus	31	1972-2002	0.18380	-0.6930
France	31	1972-2002	0.02550	-0.0197
Mauritius	31	1972-2002	0.79870	-0.0197
Average			0.26446	0.11855714
Variance			0.067371	0.200422906
Standard deviation			0.259559	0.447686169

TABLE C-3 "Autocratic" Countries (Polity Scores -10 to -6)—
Correlations with FH Scores

Country	Number of Observations	Years	Correlation Coefficient	Correlation for First Differences
China	31	1972-2002	0.34910	-0.0113
Burma	31	1972-2002	0.53420	0.3216
USSR	20	1972-1991	0.84570	0.7358
Bahrain	31	1972-2002	-0.12800	0.3623
Iraq	31	1972-2002	-0.06930	0.4152
Kuwait	30	1972-1989; 1991-2002	0.39630	0.8575
Morocco	31	1972-2002	-0.18060	0.4120
Oman	31	1972-2002	0.57210	-0.0496
Syria	31	1972-2002	-0.25580	0.0000
Yemen South	18	1972-1989	-0.78260	-0.4123
Eritrea	10	1993-2002	0.77170	0.3500
Mauritania	31	1972-2002	0.33160	-0.0197
Swaziland	31	1972-2002	0.78380	0.8647
Congo Kinshasa	20	1972-1991	0.66670	0.3294
Average			0.27392	0.296829
Variance			0.234796	0.136285
Standard deviation			0.484558	0.369168

WORKSHOP AGENDA AND PARTICIPANTS

Democracy Indicators for Democracy Assistance

January 26-27, 2007

Boston University

AGENDA

Friday, January 26, 2007

- 1:00 p.m. Meeting begins
- Opening remarks
 - Introductions
 - Brief project overview
 - Plan for the meeting
- 1:30 p.m. **Overview: USAID and Democracy Assistance Work**
- History of USAID Indicator Work
David Black, USAID
- Applicability to USAID Programming and Evaluation
Margaret Sarles, USAID
- 2:00 p.m. **Extant Indicators.** How good are they? To what degree do they fulfill USAID's objectives, and to what extent do they fall short? Particular focus on Polity, Freedom House (with its newly released subcomponents), and the new (somewhat disaggregated) index from the Economist Intelligence Unit.
- 3:00 p.m. Break
- 3:15 p.m. **Defining and Measuring Democracy.** What is democracy? Can its dimensions and subcomponents be specified? What are the boundaries of what we choose to measure? Which important aspects of society (i.e. human rights, economic freedoms, and perhaps some things labeled governance) should fall outside our definition of democracy? Should the project also include aspects of governance that do not fall within the rubric of democracy (*tout court*)?
- 6:30 p.m. Meeting Adjourns
- 7:00 p.m. Committee Working Dinner

Saturday, January 27, 2007

- 10:00 a.m. **The Aggregation Problem.** Can aggregation rules be arrived at (a) within dimensions and (b) across dimensions? Can we provide some guidance to USAID on how to define “Big-D” democracy? Or is it advisable to avoid this highest level of aggregation?
- 11:00 a.m. **History.** How important is the historical aspect of the index? What would have to be sacrificed from the current index in order for it to be extended back to 1960, 1900, or 1800?
- 11:30 a.m. **Management and Payoff.** How to make this project work? Will the necessary data be available? How big a project is this, really? How much time would it take? How much money would it cost? How would it be organized? (Should we rely primarily on students or expert staff? If the latter, would they need to be paid, and if so how much?) What is the potential payoff of this project? Is it worth the money it would take?
- 12:00 p.m. **General Discussion** (Lunch meeting). Revisit all issues to see what points of consensus have been reached and what points of disagreement remain. Try to resolve the latter. Return to issues that need more discussion.
- 1:30 p.m. **Final Recommendations and Conclusions**
- 2:00 p.m. Meeting Adjourns

PARTICIPANTS

David Black
U.S. Agency for International
Development

Michael Coppedge
Notre Dame University

John Gerring
Boston University

Andrew Green
Georgetown University

Rita Guenther
National Academies

Jo Husbands
National Academies

Gerardo Munck
University of Southern California

Margaret Sarles
U.S. Agency for International
Development

Frederic Schaffer
Harvard University

Richard Snyder
Brown University

Paul Stern
National Academies

Nicolas van de Walle
Cornell University

OUTLINE FOR A DISAGGREGATED MESO- LEVEL DEMOCRACY INDEX

John Gerring

Chapter 3 introduced the Committee's proposal to develop a disaggregated index, which we believe will better serve USAID's needs for strategic assessment and tracking. At the *meso* level, we identified 13 dimensions of democracy that may be independently assessed:

1. **National Sovereignty:** Is the nation sovereign?
2. **Civil Liberty:** Do citizens enjoy civil liberty in matters pertaining to politics?
3. **Popular Sovereignty:** Are elected officials sovereign relative to non-elected elites?
4. **Transparency:** How transparent is the political system?
5. **Judicial Independence:** How independent, clean, and empowered is the judiciary?
6. **Checks on the Executive:** Are there effective checks on the executive?
7. **Election Participation:** Is electoral participation unconstrained and extensive?
8. **Election Administration:** Is the administration of elections fair?
9. **Election Results:** Do results of an election indicate that a democratic process has occurred?
10. **Leadership Turnover:** Is there regular turnover in the top political leadership?
11. **Civil Society:** Is civil society dynamic, independent, and politically active?
12. **Political Parties:** Are political parties well institutionalized?
13. **Subnational Democracy:** How decentralized is political power and how democratic is politics at subnational levels?

The rest of this section of Appendix C elaborates on some of the issues related to the proposed index, concluding with a more detailed listing of the 13 dimensions listed above.

Components

Each dimension has multiple components, chosen with five criteria in mind: (a) centrality to the dimension, (b) centrality to the overall concept of democracy (defined minimally and maximally, as explained in the text), (c) the possible incorporation of existing data, (d) measurement precision, (e) accuracy (reliability), and (f) nonredundancy. Each component is stated in the form of a question or statement that may be coded numerically for

a given country or territory during a given year. Further work will be required in order to specify what these scales mean in the context of each question. The devil is always in the details.

Coding categories are *dichotomous* (yes/no), *categorical* (unranked), *nominal* (ranked), or *interval*. In certain cases, it may be possible to combine separate components into more aggregated nominal scales without losing information (Coppedge and Reinicke 1990). This is possible, evidently, only when the underlying data of interest are, in fact, nominal.

There are roughly 100 components in the index as currently constructed. While this may seem like quite a few, the reader is urged to consider that most of these questions—indeed, the vast majority—are very simple to answer. Thus, it should not take a country expert (or well-coached student assistant) very long to complete the questionnaire. Indeed, this is precisely the point. A longer set of questions is sometimes quicker to complete than a much shorter set of questions, if the latter are vague and ambiguous (due, we suppose, to a high level of aggregation).

For each datum, one should record (a) the coding (numerical or natural language), (b) the source(s) on which the coding was based, (c) the coder(s), (d) any revisions to the initial coding that may have been made in previous iterations of the dataset, (e) any further explanation that might be helpful, and (f) estimates of uncertainty (discussed below). Evidently, it is important that the data-storage software be capable of handling numerical and narrative responses (e.g., MS Access).

Objective/Subjective Measures

With respect to attaining greater accuracy, “hard” or “objective” indicators—based on what might be considered factual matters—are preferred over expert opinions. As one example, one might consider how to replace (or supplement) the opinion of country experts about how free the press is with a content analysis of major news outlets. Where the press is free, one would expect to find (a) a dispersion of views across news sources and (b) criticism of political leaders. Both signal the existence of the sort of open debate that is impossible if the press is constrained, and inevitable (one would think) if it is not.

At the same time, it is important to note that the development of an objective measure for a difficult concept such as press freedom is apt to be time-intensive and costly, and may not be possible at all for previous eras. Additionally, objective indicators are sometimes subject to the problem of “teaching to the test”; governments can attain higher scores by fulfilling some criterion that has little import for democracy.

The benefits of easy data collection thus must be balanced against the benefits of data efficiency, coverage, and conceptual validity.

Survey Research

A major question is whether to include dimensions that require public opinion surveys. The EIU index has lots of questions of this nature, for example, about how legitimate the general public views the election process. (“Democracy assessments” also rely centrally on surveys, though their purpose is usually not comparative [Beetham 2004].) We have opted to include relatively few questions of this nature because (a) it is very expensive to do this sort of public opinion polling on a regular basis and across all countries, (b) it is less useful if polling is conducted only in “problem” countries (for then there is no basis for comparison), (c) no such historical information is available, (d) polling questions tend to vary in form or format from country to country and year to year and hence may convey misleading information if used as a cross-national indicator, (e) in nondemocratic countries citizens may not feel free to speak openly, and (f) public perceptions are not the most valid test of a country’s level of democracy, even where civil liberties are ensured. (On the latter point, one might consider Mexico’s recent election, which many members of the public thought was highly flawed, but which outside observers seem to think was conducted with considerable fairness.)

Data Sources

For contemporary years, obtaining sufficient information to code each new component ought to be fairly easy. Sources such as the *Chronicle of Parliamentary Elections [and Developments]*, *Keesing’s Contemporary Archives*, the *Journal of Democracy* (“Election Watch”), *El Pais* (www.elpais.es), the *Statesman’s Yearbook*, *Europa Yearbook*, *Political Handbook of the World*, reports of the Inter-Parliamentary Union, the ACE Electoral Knowledge Network, *Elections Around the World* (www.electionworld.org), the International Foundation for Election Systems (www.IFES.org), the Commonwealth Election Law and Observer Group (www.thecommonwealth.org), the OSCE Office for Democratic Institutions and Human Rights (www.osce.org/odihr), the Carter Center (www.cartercenter.org), the International Republican Institute (www.iri.org), the National Democratic Institute (www.ndi.org), the Organization for American States (www.oas.org), country narratives from the annual Freedom House surveys, newspaper reports, and secondary accounts (according to subject and time period) will be invaluable. Given the project’s broad theoretical scope and empirical reach, evidence-gathering approaches must be eclectic. Multiple sources will be employed wherever possible in order to cross-validate the accuracy of underlying data.

Uncertainty

It is vital to include not only an estimate of a country's level of democracy across various dimensions and components but also a level of *uncertainty* associated with each estimate. This may be arrived at by combining two features of the analysis (a) intercoder reliability (if available) and (b) subjective uncertainty (the coder's estimate of how accurate a given score might be). Uncertainty estimates serve several functions: Scholars may include these estimates as a formal component of their analyses; they provide a signal to policymakers of where the democracy index is most (and least) assured; and they focus attention on ways in which future iterations of the index may be improved.

Finally, uncertainty estimates allow for the inclusion of countries and time periods with vastly different quantities and qualities of data—without compromising the legitimacy of the overall project. As noted, contemporary codings are likely to be associated with lower levels of uncertainty than the analogous historical codings, and countries about which much is known (e.g., France) will be associated with lower levels of uncertainty than countries about which very little is known (e.g., Central African Republic). Without corresponding estimates of uncertainty, an index becomes hostage to its weakest links; critics gravitate quickly to countries and time periods that are highly suspect, and the validity of the index comes under harsh assault—even if the quality of other data points is more secure. With the systematic use of uncertainty estimates, these very real difficulties are brought directly into view by granting them a formal status. In so doing, the legitimacy of the larger enterprise is enhanced, and misuses are discouraged.

Time

The dataset is assumed to be annual, though it might be coded at longer intervals in earlier historical periods. (One minor question to consider is whether codings should refer to the state of affairs pertaining at the end of the designated period (December 31), or to a mean value across the period of observation [January 1–December 31].)

It is strongly urged that the index—or at least some elements of it—be extended back in time, preferably to 1800. There are several reasons for this. First, if one wishes to judge trends, a trend line is necessary. And the longer the trend line, the more information will be available for analysis. Consider the question of how Ukraine is doing now—for example, in 2008. If a new index provides data only for that year, or several years prior, the meaning of a “5” (on some imagined scale) is difficult to assess. Similarly, a purely contemporary index is unable to evaluate the question of democratic “waves” occurring at distinct points in historical time

(Huntington 1991) or of distinctive “sequences” in the transition process (McFaul 2005). If we wish to judge the accuracy of these hypotheses (and many others) we must have at our disposal a substantial slice of historical time.

Second, insofar as we wish to understand causal relations—what causes democracy and what democracy causes—it is vital to have a long time series so that causes and effects can be effectively disentangled. (Of course, this does not assure that they will be disentangled; but with observational data it is virtually a prerequisite.)

Third, recent work has raised the possibility that democracy’s effects are long term, rather than (or in addition to) short term (Gerring et al 2005, Converse and Kapstein 2006, Persson and Tabellini 2006). Indeed, it is quite possible that the short-term and long-term effects of democracy are quite different (plausibly, long-term effects are more consistent, and more positive along various developmental outcomes, than short-term effects). Consideration of these questions demands a historical coding of the key variable.

For all these reasons, we think it unlikely that any new index would displace Freedom House, Polity, and ACLP unless it can match the historical coverage of these well-established indices.

Summary Scores

For each dimension, a summary score will be suggested. Evidently, this task of aggregation is devilish, for all the reasons just reviewed. Yet, it should be considerably easier to solve at this level than at the level of Big D democracy. Thus, we propose to aggregate the results for each component so as to arrive at a single score for each of the 13 dimensions. This score will be expressed on a scale from 1 to 10, providing a snapshot view of how each country, in a given year, performs on that dimension.

We feel confident that, with the aid of the underlying components listed in the index below, it will be possible for those knowledgeable about a country to reach agreement on the (approximate) level of national sovereignty, popular sovereignty, and so on enjoyed by that country in a given year. A country’s score along these 13 dimensions comprises its *Democracy Profile*.

This level of aggregation seems feasible, and should be easy to compare across countries and through time. We also believe that this is a useful level of aggregation. It says something meaningful, something that should be understandable to all observers. It will allow USAID and other international actors a way of gauging progress and regress; it may even provide a way of gauging the relative success of different programs—though problems of causal attribution are inevitably knotty.

We are considerably less confident that it will be possible to reach agreement in aggregating *across* the 13 dimensions to reach a single, summary score for each country in a given year—“Big-D” democracy.

Logistics

In order to manage a project of this scope without losing touch with the particularities of each case, it is necessary to marry the virtues of cross-national data with the virtues of regional expertise. As currently envisioned, the project relies primarily upon country experts to do the case-by-case coding. Student assistants may be employed in a supporting role (e.g., to fetch data). These coding decisions will be supervised by several regional experts who are permanently attached to the project and who will work to ensure that coding procedures across countries, regions, and time periods are consistent. Extensive discussion and cross-validation will be conducted at all levels, including intercoder reliability tests.

We strongly advise an open and transparent system of commentary on the scores that are proposed for each country, after initial questionnaires are completed by country experts but before results are finalized. This might include a Web-based Wikipedia-style discussion in which interested individuals are encouraged to comment on the scores provisionally assigned to the country or countries that they know well. This commentary might take the form of additional information—perhaps unknown to the country expert—that speaks to the viability of the coding. Or it might take the form of extended discussions about how a particular question applies to the circumstances of that country. Naturally, some cranky participants may be anticipated in such a process. However, the Wikipedia experience suggests that there are many civic-minded individuals, some of them quite sophisticated, who may be interested in engaging in this process and may have a lot to add. At the very least, it may provide further information upon which to base estimates of uncertainty (as discussed above). Final decisions, in any case, would be left to a larger committee.

Evidently, different components will involve different sorts of judgments and different levels of difficulty. Some issues are harder than others, and will require more codings and recodings. As a general principle, wherever low intercoder reliability persists for a given question, that question should be reexamined and, if possible, reformulated.

It is important that the process of revision be *continual*. Even after the completed dataset is posted, users should be encouraged to contribute suggestions for revision and these suggestions should be systematically reviewed.

Pilot Tests

Before USAID, or any agency, undertakes a commitment to develop—and maintain—a new democracy index, it is important that it be confident of the yield. Thus, we recommend several interim tests of a “pilot” nature.

One of the principal claims of this index is that greater intercoder reliability will be achieved when the concept of democracy is disaggregated. This claim may be probed through intercoder reliability tests *across* the leading democracy indices. A pilot test of this nature might be conducted in the following manner: Train the same set of coders to code all countries (or a subset of countries) in a given year according to guidelines provided by Freedom House, Polity, and the present index. Each country-year would receive several codings by different coders, thus providing the basis for an intercoder reliability test. These would then be compared across indices. Since the coders would remain the same, varying levels of intercoder reliability should be illustrative of basic differences in the performance of the indices. Of course, there are certain methodological obstacles to any study of this sort. One must decide how much training to provide to the coders, and how much time to give them. One must decide whether to employ a few coders to cover all countries, or have separate coders for each country. One must decide whether to hire “naïve” coders (e.g., students) or coders well versed in the countries and regions they are assigned to code (the “country expert” model). In any case, we think the exercise worthwhile, not only because it provides an initial test of the present index but also because it may bring a level of rigor to a topic—political indicators—that has languished for many years in a highly unsatisfactory state.

THE INDEX

Dimensions

1. **National Sovereignty:** Is the nation sovereign?
2. **Civil Liberty:** Do citizens enjoy civil liberty in matters pertaining to politics?
3. **Popular Sovereignty:** Are elected officials sovereign relative to non-elected elites?
4. **Transparency:** How transparent is the political system?
5. **Judicial Independence:** How independent, clean, and empowered is the judiciary?
6. **Checks on the Executive:** Are there effective checks on the executive?

7. **Election Participation:** Is electoral participation unconstrained and extensive?
8. **Election Administration:** Is the administration of elections fair?
9. **Election Results:** Do results of an election indicate that a democratic process has occurred?
10. **Leadership Turnover:** Is there regular turnover in the top political leadership?
11. **Civil Society:** Is civil society dynamic, independent, and politically active?
12. **Political Parties:** Are political parties well institutionalized?
13. **Subnational Democracy:** How decentralized is political power and how democratic is politics at subnational levels?

Clarifications

“Party” may refer to a longstanding coalition such as the CDU/CSU in Germany if that coalition functions in most respects like a single party. The identity of the party may be obscured by name changes. (If the party/coalition changes names but retains key personnel and is still run by and for the same constituency then it should be considered the same organization.)

“Executive” refers to the most powerful elective office in a country (if there is one)—usually a president or prime minister.

Wherever there is disparity between formal rules (constitutional or statutory) and actual practice, coding decisions should be based on the latter.

Unless otherwise specified, the geographic unit of analysis is the (sovereign or semi-sovereign) nation-state. Evidently, there is enormous heterogeneity within large nation-states, necessitating judgments about which level of coding corresponds most closely to the mean value within that unit. Where extreme heterogeneity exists vis-à-vis the variable of interest it may be important to include a companion variable that would indicate high within-country variance on that particular component. One thinks of contemporary Sri Lanka and Colombia—states where the quality of democracy is quite different across regions of the country.

Questions pertaining to elections may be disaggregated according to whether they refer to elections for the (a) lower house, (b) upper house, or (c) presidency. In some cases, (b) and/or (c) is nonexistent or inconsequential, in which case it should be ignored. If no election occurs in a given year, then many of these questions should be left unanswered (unless of course rules or norms pertaining to elections have changed in the interim). If more than one election occurs in a given year there will be two entries for that country in that year. (This complicates data

analysis, but it is essential to the purpose of the dataset, which is to provide primary-level data that can be used for further analysis.)

At some point, coding responses must be added to this questionnaire. Such responses may be dichotomous, multichotomous, or continuous, depending upon the question. However, we suggest that all original coding scales (where coding decisions are required) be comprised of no more than five categories. A larger number of options may create greater ambiguity. In any case, these response options should be as operational as possible. It should be clear what a “3” means with respect to the question at hand.

1. *National Sovereignty*

General question: Is the nation sovereign?

Is the territory independent of foreign domination? (Note: We are not concerned here with pressures that all states are subject to as part of the international system.)

2. *Civil Liberty*

General questions: Do citizens enjoy civil liberty in matters pertaining to politics?

Note: Civil liberties issues pertaining specifically to elections are covered in later sections.

Does the government directly or indirectly attempt to censor the major media (print, broadcast, Internet)? Indirect forms of censorship might include politically motivated awarding of broadcast frequencies, withdrawal of financial support, influence over printing facilities and distribution networks, selective distribution of advertising, onerous registration requirements, prohibitive tariffs, and bribery. (See recent index of Internet freedom developed by the Berkman Center for Internet and Society, Harvard University.)

Of the major media outlets, how many routinely criticize the government?

Are individual journalists harassed—i.e., threatened with libel, arrested, imprisoned, beaten, or killed—by government or nongovernmental actors while engaged in legitimate journalistic activities?

Is there self-censorship among journalists when reporting on politically sensitive issues?

Are works of literature, art, music, and other forms of cultural expression censored or banned for political purposes?

Do citizens feel safe enough to speak freely about political subjects in their homes and in public spaces?

Is it possible to form civic associations, including those with a critical view of government?

Is physical violence (e.g., torture) and/or arbitrary arrest targeted at presumed opponents of the government widespread?

Are certain groups systematically discriminated against by virtue of their race, ethnicity, language, caste, or culture to the point where it impairs their ability to participate in politics on an equal footing with other groups? (*Note:* This question pertains to citizens only [not non-citizens] and does not cover issues of disenfranchisement, which are included in a later section.)

If so, how large (as a percentage of the total population) is this group(s)?

3. *Popular Sovereignty*

General question: Are elected officials sovereign relative to nonelected elites?

Are there national-level elections (even if only pro forma)?

If yes, are the governments that result from these elections fully sovereign—in practice, not merely in constitutional form—vis-à-vis any nonelective bodies whose members are not chosen by, or removable by, elected authorities (e.g., a monarchy, the military, and the church)? Note that this does not preclude extensive delegation of authority to nonelective bodies such as central banks and other agencies. But it does presume that the members of these nonelective authorities are chosen by, and may be removed, in circumstances of extreme malfeasance, by elective authorities. This power of removal must be real, not merely formal. Thus, while constitutions generally grant power to civilian authorities to remove military rulers, it is understood that in some countries, during some periods, an action of this nature would not be tolerated. In most cases, it will be clear to those familiar with the countries in question when this sort of situation obtains, though there may be questions about the precise dates of transition (e.g., when Chilean political leaders regained control over the military after the Pinochet dictatorship).

4. *Transparency*

General question: How transparent is the political system?

Note: this section pertains to the polity as a whole, while some other questions listed below pertain to particular sections of the polity (e.g., election administration).

Are government decisions made public in a timely fashion and otherwise made accessible to citizens?

Are decision-making processes open to public scrutiny, for example, through committee hearings?

5. *Judicial Independence*

General question: How independent, clean, and empowered is the judiciary?

Is the judiciary independent of partisan-political pressures?

Is the judiciary noncorrupt?

Is the judiciary sufficiently empowered to enforce the laws of the land, including those pertaining to the ruling elite (or is its power so reduced that it cannot serve as a check on other branches of government)?

6. *Checks on the Executive*

General question: Are there effective checks—other than elections—on the exercise of power by the executive?

Note: Questions pertaining to electoral accountability are addressed elsewhere.

Constitutionality

Does the executive behave in a constitutional manner (i.e., according to written constitutional rules or well-established constitutional principles)?

Term limits

If the executive is elected directly by the general electorate (or through an electoral college), are there term limits?

If so, what are they?

Are they respected (at this point in time)?

The legislature

Is the executive able to control the legislature by undemocratic means (e.g., by manipulating legislative elections, by proroguing the legislature, by buying votes in the legislature)?

Is the executive able to make major policy decisions without legislative approval, i.e., without passing laws? Can the executive rule by fiat?

The judiciary

Is the executive accountable to the judiciary—which is to say, is the judiciary prepared to enforce the constitution, even when in conflict with the executive?

7. *Election Participation*

General question: Is electoral participation unconstrained and extensive?

Suffrage

What percent of citizens (if any) are subject to de jure and de facto eligibility restrictions based on ascriptive characteristics other than age (e.g., race, ethnicity, religion)?

What percent of the population are excluded from suffrage by virtue of being permanent residents (noncitizens)?

Turnout

Note: This variable is meaningless in the absence of free and fair elections. Therefore, although data may be collected for all countries, it should be considered an aspect of democracy only where countries score above some minimal level on Election Administration.

What percent of the adult (as defined by the country's laws) electorate turned out to vote?

8. *Election Administration*¹

General question: Is the administration of elections fair?

Election law

At this time, are regularly scheduled elections—past and future—on course, as stipulated by election law or well-established precedent? (If the answer is no, the implication is that they have been suspended or postponed in violation of election law or well-established precedent.)

Are there clear and explicit sets of rules for the conduct of elections and are the rules clearly disseminated (at the very least, to political elites in the opposition)?

Election commission

Note: Election commission refers to whatever government bureau(s) is assigned responsibility for setting up and overseeing elections.

Is it unbiased and independent of partisan pressures *or* balanced in its representation of different partisans?

Does it have sufficient power and/or prestige to enforce its own provisions? (Are its decisions respected and carried out?)

Registration

Are electoral rolls updated regularly?

Do they accurately reflect who has registered? (If the election rolls are not made public, then the answer is assumed to be No.)

Do names of those registered appear on the rolls at their local polling station (as they ought to)?

Integrity of the vote

Are all viable political parties and candidates granted access to the ballot (without unduly burdensome qualification requirements)?

Are opposition candidates/parties subject to harassment (e.g., selective prosecution, intimidation)?

Is the election process manipulated through other means (e.g., changing age or citizenship laws to restrict opposition candidate's access to the ballot, stalking horse candidates, snap elections scheduled without sufficient time for the opposition to organize)?

Are election choices secret (or are there violations)?

¹This section draws on Munck (2006).

Is vote-buying (bribery) and/or intimidation of voters widespread?

Are other forms of vote fraud (e.g., ballot-stuffing, misreporting of votes) widespread?

What percent of polling stations did not open on time, experienced an interruption, ran out of voting materials, or experienced some other sort of irregularity?

What was the percentage of lost or spoiled ballots?

Media

Do all parties and candidates have equal access to the media? Equal access is understood as (a) all candidates or parties for a particular office are treated equally (thus granting an advantage to small parties or minor candidates) or (b) access to the media is in rough proportion to the demonstrated support of a party or candidate in the electorate.

Is election reportage (reportage about politics during election periods) biased against certain parties and/or candidates?

Campaign finance

Are there disclosure requirements for large donations?

If so, are these effective (i.e., are they generally observed)?

Is public financing available?

If so, does it constitute at least one-third of the estimated expenditures by candidates and/or parties during the course of a typical campaign?

Does the incumbent enjoy unfair advantages in raising money by virtue of occupying public office? Unfair advantage involves such things as (a) a levy on civil servants to finance the party's campaigns, (b) widespread and organized use of civil servants for campaign purposes, or (c) use of government materiel for campaign purposes.

Is campaign spending heavily tilted in favor of the incumbent party or candidate(s)?

That is, does the incumbent party or candidate(s) expend more financial resources than their support in the electorate (as judged by polls or general impressions) or the legislature would indicate?

Note: Where campaign expenditures are unreported, or such reports are unreliable, they may be estimated from each party's campaign activity, e.g., number of political advertisements on TV, radio, or billboards.

Election monitors

Were election monitors from all parties and/or from abroad allowed to monitor the vote at polling stations across the country?

How many polling stations (percent) were attended by election monitors (other than those representing the ruling party or clique)?

9. Election Results

General question: Do results of an election indicate that a democratic process has occurred?

What percent of the vote was received by the largest party or winning candidate in the final (or only) round?

Specify name of party or candidate:

What percent of the vote was received by the second largest party or second most successful candidate in the final round?

Specify name of party or candidate:

What percent of the seats in the lower/upper house was obtained by the largest party?

Specify name of party:

What percent of the seats in the lower/upper house was obtained by the second largest party?

Specify name of party:

Do the official results conform, more or less, to actual ballots cast (as near as that can be estimated)?

What was the general verdict by international election monitors and or the international press vis-à-vis the democratic quality of this election, i.e., how fair was it?

Note: If there was disagreement, then please report the mean (average) result, weighting each group by its level of involvement in overseeing this election.

Did losing parties/candidates accept the essential fairness of the process and the result?

10. *Leadership Turnover*

General question: Is there regular turnover in the top political leadership?

Note: Turnover may be regarded as a sufficient condition of effective electoral competition. If turnover occurs (by democratic instruments), contestation must be present—though it may of course still be flawed.

Executive

How many years has the current executive been in office? (*Source:* "YRSOFFC" variable from the DPI.)

How many consecutive terms has the current executive served?

Did the last turnover in power occur through democratic means (e.g., an election, a loss of confidence in the legislature, or a leader's loss of confidence in his/her own party)?

Ruling party/coalition

How many years has the current ruling party or coalition been in office? (*Source:* "PRTYIN" variable from the DPI.)

How many consecutive terms has the current ruling party or coalition served?

Note: relevant only where elections fill the major offices.

Did the last turnover in power occur through democratic means (e.g.,

an election, a loss of confidence in the legislature, or a leader's loss of confidence in his/her own party)?

11. *Civil Society*

General question: Is civil society dynamic, independent, politically active, and supportive of democracy?

Notes:

a. "Civil society organization" refers to any of the following: an interest group, a social movement, church group, or classic NGO, but *not* a private business, political party, or government agency. Must be at least nominally independent of government and the private sector.

b. Questions about civil liberties, of obvious significance to civil society, are covered in a separate section.

Existing indicators: the Civil Society Index compiled by the Global Civil Society Project.

How much support for democracy is there among citizens of the country? (*Sources:* World Values Surveys, Eurobarometer, Afrobarometer, Latinobarometer [see EIU].)

What is the level of literacy (a presumed condition of effective participation)? (*Source:* WDI.)

What percent of citizens regularly listen to or read the national news?

Are civil society organizations generally independent of direct government influence (or are they manipulated by the government and its allies such that they do not exercise an independent voice)?

Are there any sizeable civil society organizations that are routinely critical of the government?

Are major civil society organizations—representing key constituencies on an issue—routinely consulted by policymakers on policies relevant to their members (e.g., by giving testimony before legislative committees)?

12. *Political Parties*

General question: Are political parties well institutionalized?

Notes:

a. Questions about the freedom to form parties and participate in elections are included under Election Administration.

b. Questions below refer to all parties in a polity, considered as a whole. However, larger parties should be given greater weight in calculating answers so that the party system is adequately represented.

Are there well-understood rules governing each party's business and, if so, are these rules generally followed?

Is there a clearly identifiable group of party members and is this group relatively stable from year to year?

Do parties issue detailed policy platforms (manifestos)?

Do parties hold regular conventions and, if so, are these conventions sovereign (in the sense of making final decisions on party polity and procedure)?

Do parties have local sections (constituency groups), or are they centered on the capital and on a restricted group of local notables?

13. *Subnational Government*

General question: How democratic is politics at subnational levels?

Note: "Subnational government" refers to governments at regional and local levels.

How centralized is power within the polity, taking all factors into account (for a useful discussion of various relevant factors see Rodden 2004)? As a way of calibrating this, Switzerland may be said to define the decentralized extreme while New Zealand defines the centralized extreme among democratic polities. Most authoritarian regimes are highly centralized, but not all (e.g., failed states such as Afghanistan or Somalia). To clarify, the question refers to the *relative* power balance between national and subnational levels; it does not attempt to judge the actual strength of control at either level. That is, whether both levels of government are weak or strong is irrelevant; what is relevant is only their power relative to each other. The question pertains to practical power not to formal/constitutional power. Note that centralization is usually not considered a definitional component of democracy: New Zealand, most would agree, is no less democratic than Switzerland. However, if power is highly centralized in a very large country—say, India—one may infer a significant problem of local accountability. In any case, the degree of centralization/decentralization gives meaning to the next question.

How democratic are electoral politics at the subnational level? If practices differ appreciably between national and subnational levels, and perhaps even between regional and local levels, it may be necessary to complete the previous sections—Election Participation, Election Administration, Election Results, Leadership Turnover—for different levels of government.

References

- Converse, N., and Kapstein, E.B. 2006. "The Economics of Young Democracies: Policies and Performance." Working Paper No. 85, Center for Global Development (March).
- Coppedge, M., and Reinicke, E.B. 1990. "Measuring Polyarchy." *Studies in Comparative International Development* 25:51-72.
- Europa Yearbook. [various years] *The Europa Yearbook*. London: Europa Publications.

- Gerring, J.; Bond, P.; Barndt, W.; and Moreno, C. 2005. Democracy and Growth: A Historical Perspective. *World Politics* 57(3):323-364.
- Huntington, Samuel P. 1991. *The Third Wave: Democratization in the Late Twentieth Century*. Norman, OK: University of Oklahoma Press.
- McFaul, M. 2005. Transitions from Postcommunism. *Journal of Democracy* 16(3): 5-19.
- Munck, G. L. 2006. Standards for Evaluating Electoral Processes by OAS Election Observation Missions. Paper prepared for Organization of American States.
- Persson, T., and Tabellini, G. 2006. Democratic Capital: The Nexus of Political and Economic Change. NBER Working Paper. No. 12175.
- Rodden, J. 2004. Comparative Federalism and Decentralization: On Meaning and Measurement. *Comparative Politics* (July):481-500.

D

Understanding Democratic Transitions and Consolidation from Case Studies: Lessons for Democracy Assistance

*Co-sponsored by The National Academies
and the Center on Democracy, Development, and the Rule of Law,
Stanford University*

AGENDA

March 5-6, 2007

Sunday, March 4, 2007

7:30 p.m. **Planning Meeting and Working Dinner**

Monday, March 5, 2007

9:00 **Welcome Remarks**

9:05 **Opening Address and Overview**

*Michael McFaul, Director, CDDRL, Stanford University
Jack Goldstone, George Mason University and Chair, NAS-
CEUDAP Committee*

9:35 **Panel I: What Have We Learned About Democratic
Transitions: Pacts or Protests?**

Moderator: Jack Goldstone, George Mason University

Nancy Bermeo, Princeton University
Adrian Karatnycky, Freedom House, Inc.
Terry Karl, Stanford University
Michael McFaul, Stanford University

10:20-10:40 **Break**

10:40-11:00 **Panel Discussion Session**

11:00-12:00 **Open Discussion Session**

12:00 **LUNCH**

1:00 **Panel II: What Have We Learned About Democratic
 Transitions: Are Certain Socioeconomic or Political
 Conditions Required?**

Moderator: Larry Garber, New Israel Fund

Sheri Berman, Barnard College
Michael Bratton, Michigan State University
Jason Brownlee, University of Texas at Austin
Lucan Way, University of Toronto

1:45-2:05 **Panel Discussion Sessions**

2:05-2:45 **Open Discussion Sessions**

2:45-3:00 **Break**

3:00 **Panel III: What Have We Learned About Democratic
 Transitions: What Comes First? What Role Will Foreign
 Assistance Play?**

Moderator: Jeremy Weinstein, Stanford University

Tom Carothers, Carnegie Endowment for International Peace
Gerry Hyman, Center for Strategic and International Studies
Cynthia McClintock, George Washington University
Risto Volanen, State Secretary, Finish Prime Minister's Office

3:45-4:05 **Panel Discussion Session**

4:05-4:45 **Open Discussion Session**

6:30 **CONFERENCE DINNER**

Tuesday, March 6, 2007

- 9:00 **Panel IV: What Have We Learned About Democratic Consolidation: Do Certain Rules and Procedures Work Better Than Others, and Can They Be Fitted to Known Conditions?**
Moderator: John Gerring, Boston University
Gerardo Munck, University of Southern California
Marc Plattner, National Endowment for Democracy
Andy Reynolds, University of North Carolina at Chapel Hill
Andreas Schedler, Centro de Investigación y Docencia Económicas, Mexico City
- 9:45-10:05 **Panel Discussion Session**
- 10:05-10:45 **Open Discussion Session**
- 10:45 **Break**
- 11:00 **Panel V: What Have We Learned About Democratic Consolidation: Can We Combine Democracy Assistance and Other Forms of Aid to Promote Consolidation?**
Moderator: Mitch Seligson, Vanderbilt University
Larry Diamond, Stanford University
Amichai Magen, Stanford University
Philippe Schmitter, European University Institute, Florence
- 11:45-12:05 **Panel Discussion Session**
- 12:05-12:45 **Open Discussion Session**
- 12:45 **LUNCH**
- 2:00 **Conference Roundtable I**
Democracy Promotion: Developing Guidelines for Foreign Assistance
Moderator: Kathryn Stoner-Weiss, Stanford University
- 3:00 **Conference Roundtable II**
Democracy Consolidation: Developing Guidelines for Foreign Assistance
Moderator: Jack Goldstone, George Mason University
- 4:00 **Concluding Remarks**
Michael McFaul, Stanford University
Jack Goldstone, George Mason University
- 4:30 **Meeting Adjourns**

PARTICIPANTS

Tabitha Benney
National Academies

Sheri Berman
Barnard College

Nancy Bermeo
Princeton University

Michael Bratton
Michigan State University

Jason Brownlee
University of Texas at Austin

Tom Carothers
Carnegie Endowment for
International Peace

Larry Diamond
Stanford University

Jim Fearon
Stanford University

Larry Garber
New Israel Fund

John Gerring
Boston University

Jack A. Goldstone
George Mason University

Rita Guenther
National Academies

Jo Husbands
National Academies

Gerry Hyman
Center for Strategic and
International Studies

Adrian Karatnycky
Freedom House, Inc.

Terry Karl
Stanford University

Abe Lowenthal
University of Southern California

Amichai Magen
Stanford University

Cynthia McClintock
George Washington University

Michael McFaul
Stanford University

Gerardo Munck
University of Southern California

Marc Plattner
National Endowment for
Democracy

Andy Reynolds
University of North Carolina at
Chapel Hill

Andreas Schedler
Facultad Latinoamericana de
Ciencias Sociales

Philippe Schmitter
European University Institute,
Florence

Mitchell Seligson
Vanderbilt University

Kathryn Stoner-Weiss
Stanford University

Risto Volanen
Office of the Prime Minister

Lucan Way
University of Toronto

Jeremy Weinstein
Stanford University

Jennifer Windsor
Freedom House

E

Field Visit Summary Report¹

OVERVIEW OF NATIONAL ACADEMIES' MISSION AND TASKS

The field visits were part of a larger project conducted by the National Academies (NA) for the U.S. Agency for International Development (USAID), the purpose of which was to develop an overall research and analytic design that will lead to specific findings and recommendations for the Strategic and Operational Research Agenda (SORA) of the democracy and governance (DG) programs. These findings and recommendations were developed through the vetting of a variety of methodologies for assessing and evaluating democracy assistance programs.

OBJECTIVES OF FIELD VISITS

In support of these overall project objectives, the field visits were intended to serve two major purposes:

1. The collection of information for the NA committee to inform its recommendations, in particular to increase members' understanding of:

- how USAID programs are developed and implemented in the field as background for its recommendations to improve program evaluation and understanding of program successes and failures,
- what data, evidence, and other resources are primarily or

¹Some of the material in this Appendix also appears in Chapters 6 and 7.

uniquely available in the mission or in country to support improved program evaluation,

- the perspectives of mission personnel and USAID implementers regarding the feasibility of potential options for improving program evaluation;

2. to provide an opportunity to explore a “proof of concept” of the committee’s preliminary recommendations, in particular the feasibility of introducing more rigorous approaches to program evaluation.

SELECTION OF FIELD VISIT SITES

Three countries were selected as the sites of the field visits conducted by teams of consultants and staff: Albania, Peru, and Uganda. In particular, the selection was based primarily on the stage of program development within a country’s DG portfolio, the breadth of USAID programming, and the depth of USAID programming (as determined by long-term funding in multiple program areas of interest; see “Current and Recent USAID Projects at the Time of Field Visits” at the end of this appendix for a list of the major DG projects in each country). In each country selected, the DG staff were at the stage of developing new projects, offering an optimal opportunity to explore options for program design that may be more or less suited for various research methodologies. The NA field team members (see “Consultant Biographies” at the end of this appendix) were thus able to understand a variety of projects at the stage of their inception, the point at which new methodologies would be most effectively designed to maximize confidence about the impact of projects and under what conditions. These considerations guided the selection of cases across geographically and politically distinct regions of the world (Central Europe/Post-Communist, Latin America/Post-Military Rule, Africa/Post-Conflict).

While there is no single point at which DG programs can be most effectively designed, implemented, or evaluated, the initial stages of development and design provide the most fruitful points at which innovative yet feasible options may be considered. Each field team therefore selected one or more projects and worked closely with USAID Mission DG officers, project implementers, and local partners through a series of in-depth conversations to understand the various opportunities and challenges presented by newly proposed program designs, data collection, and more rigorous evaluation techniques. A fuller discussion of these proposed program designs in each country visited follows.

KEY OBSERVATIONS AND FINDINGS FROM FIELD VISITS²

There are ample opportunities for improving the methodology of program monitoring and evaluation within the DG sector. This is in large part due to the well-developed existing USAID evaluation procedures. To maximize these opportunities, various approaches to evaluation must be selected based on program goals and program designs. This should involve the provision of assistance (e.g., visits by specialists in program monitoring and evaluation (M&E) from USAID/Washington to missions during the project conceptualization stage as well as subsequent stages of M&E development.

Improvements in program evaluation need not be expensive. Maximizing existing mechanisms (surveys and other data collection systems) and strategically targeting sample populations and control groups can result in more robust findings at a cost savings overall.

By improving program evaluation, the impact of USAID programs can be more accurately assessed and documented. Creating knowledge of program impacts through rigorous evaluation is the best way to identify and take advantage of lessons learned.

Institutional knowledge gained through these experiences should be shared within and beyond the mission to affect learning on a broader, agency-wide basis.

Building on Current Tools and Approaches

Several current practices of mission staff demonstrate the necessary willingness to maximize reasonable opportunities for learning and provide the basis for more solid inferences over time. Currently, as a part of ongoing DG programs, mission staff collect regular and systematic information about those who receive training through USAID-funded programs. This approach to data collection should be encouraged and expanded to complement other more rigorous methodologies described below.

Similarly, implementers working with USAID have developed elaborate mechanisms for quarterly data collections pertinent to their programs. To maximize the potential represented by these mechanisms, data collected should be directed toward understanding outcomes and impacts over outputs. Similarly, mechanisms created by local implementers should be strategically collected and analyzed to maximize cost benefits and

²This text is drawn from memos prepared for the committee by three of its field consultants—Thad Dunning, Yale University (Peru); Devra Cohen Moehler, Cornell University (Uganda); and Dan Posner, University of California at Los Angeles (Albania)—and reflects their judgments and assessments.

efficiencies. For example, collecting local government data in the form of smaller, cost-effective samples from municipalities would be beneficial. Furthermore, this information should be fully transferable to USAID for learning purposes. Most important, these mechanisms should be consistent with key program design elements requiring consideration at the initial stages of program development.

Measurement of Outcome Indicators

Indicators gathered in connection with past programs tend to be measures of “outputs” or very proximate outcomes. Examples of these output indicators include, in the context of a decentralization program, the number of relevant municipal officials trained by the implementer or the percent of target municipalities who agree to an assistance plan. Although these output measures may be useful and necessary for monitoring the *performance* of local implementers or to assess short-term progress on the process of implementing a program, they are less helpful for measuring the outcomes that the programs hope to promote. To improve assessment of the impact of USAID programs on ultimate objectives, it is important to gather data to the extent possible on outcome variables. One example gathered in connection with the decentralization program was the percentage of local citizens who rate the quality of local government services as “good” or “very good.”

Controls

Most program evaluations involve indicators gathered only or mostly on “treated” units (those groups, individuals, or organizations who were assisted by USAID). Sometimes this is unavoidable, as when a program works with only one unit or actor (e.g., the Congress). At other times, however, it is possible to find comparison units that would be useful for assessing the impact of U.S. interventions.

Using control groups is invaluable for attributing impact to a USAID program. For example, without a control group it is impossible to know if the change in local party development is a result of a USAID intervention or another factor such as change in national party law, economic growth, or better media coverage.

Gathering outcome measurements on control units need not be prohibitively costly. The cost of modifying the 2003 and 2005 national surveys in Peru conducted by the Latin American Public Opinion Project (LAPOP) to include a sample of residents in control group municipalities would likely have run around \$15,000 per survey, a small investment when compared to the \$20 million cost of the program over five years.

Opportunities for Randomization

Comparisons across units or groups with which USAID partners worked and those with which they did not are only partially informative about the impact of USAID interventions. For example, differences across these groups could reflect preexisting differences and unobserved confounders, rather than the impact of the intervention. Similarly, selection bias could account for the variation in performance between the treatment and control groups.

One of the ways that social scientists sometimes approach this difficulty is through random assignment of units to treatment. In the context of decentralization, for example, the municipalities with which USAID implementers work could be determined by lottery. Subsequent differences between treated and untreated municipalities are likely to be due to the intervention, since other factors will be roughly balanced across the two groups of municipalities.

Randomization is not feasible for many kinds of programs, and there can be a range of practical obstacles; yet these are also often surmountable. In addition, experimental designs need not be expensive; additional costs can be offset by savings introduced by appropriate designs.

SAMPLE PROPOSED PROGRAM EVALUATION DESIGNS FROM THREE FIELD VISITS³

Selected Designs from Albania: Rule of Law Programs

A major part of USAID's DG-related activities in Albania involved increasing the effectiveness and fairness of legal sector institutions. With one possible exception, none of these rule of law activities are amenable to randomized evaluation. This is because they each deal with either (a) technical assistance to a single unit (e.g., the Inspectorate of the High Council of Justice, the Inspectorate of the Ministry of Justice, the High Inspectorate for the Declaration and Audit of Assets, the Citizen's Advocacy Office, and the National Chamber of Advocates), (b) support for the preparation of a particular piece of legislation (e.g., the Freedom of Information Act and Administrative Procedures Code, a new conflict of interest law, and a new press law), or (c) support for a single activity, such as the implementation of an annual corruption survey. For a randomized evaluation of the efficacy of these activities to be possible they would have to be, in principle, implementable across a large number of units,

³In addition to this group of selected projects discussed here, several others were analyzed by the field teams.

which these are not. There is only one Inspectorate of the High Council of Justice, only one conflict of interest law being prepared, and only one National Chamber of Advocates being supported, so it is not possible to compare the impact of support for these activities both where they are and are not being supported, and certainly not across multiple units. The best—indeed, only—way to evaluate the success of these activities is to identify the outcomes they are designed to affect, measure these outcomes both before and after the activities have been undertaken, and compare these measures.

The trick, however, is to find appropriate measures of the outcomes that the activities are designed to affect, and this is frequently far from straightforward. For example, the goal of the technical assistance to the Inspectorates of the High Council of Justice and the Ministry of Justice is to improve the transparency and accountability of the judiciary and to increase public confidence in judicial integrity. The latter can be measured fairly easily using public opinion polls that probe respondents' trust in the judiciary and perceptions of its integrity (these would be administered before and after the period during which technical assistance was offered, and the results of the polls compared). However, measuring the degree to which the judiciary is transparent and accountable is much more difficult. Part of the problem stems from the fact that transparency and accountability can only be ascertained vis-à-vis an (unknown) set of activities that should be brought to light and an (unknown) level of malfeasance that needs to be addressed. For example, suppose that, following the implementation of the programs designed to support the Inspectorate of the High Council of Justice, we observe that three judges are brought up on charges of corruption. Should this be taken as a sign that the activities worked in generating greater accountability? Compared to a baseline of no prosecutions, the answer is probably yes, to at least some degree. But knowing just how effective the activities were depends on whether there were just three corrupt judges who should have been prosecuted or whether there were, in fact, *twenty*, in which case prosecuting the three only scratched the surface of the problem, or whether the prosecutions might be selective with the targets chosen for political reasons. Parallel problems affect other rule of law initiatives, such as efforts to improve the ability of lawyers to police themselves.

A slightly different evaluation problem arises with respect to the activities designed to support the drafting of various pieces of legislation. One fairly straightforward measure of success in this area is simply whether or not the law was actually drafted, and, if so, whether it included language that will demonstrably strengthen the rule of law. But assessing whether or not USAID's support had any impact requires weighing the counterfactual question: Would the legislation have been drafted without USAID's support

and what would it have looked like? If the answers to these questions are that the legislation would not have been drafted or that the language in the resulting law would not have been optimal, then we can judge the support from USAID to have been successful to the extent that the result we observe is better than this counterfactual outcome. The broader problem, however, is that achieving the overarching strategic objective of strengthening the rule of law will involve more than just getting legislation drafted but also getting it passed and then having it enforced. The point is that the measurable outcome of the USAID-sponsored activity is several steps removed from the true goals of the intervention, and any assessment of “success” in these areas must be interpreted in this light. This is equally true with respect to other activities, such as technical assistance to aid the Albanian government in the establishment of a copyright office or an office of patents and trademarks. Whether these institutions, once created, will have any impact on protecting intellectual property will depend on much more than whether or not a formal office designed to do so has been established.

The larger point that this discussion hints at is that many of the activities in the rule of law area involve the creation of laws or the strengthening of institutions whose existence is a prerequisite for a legal system that works, and that supports democracy and market reform. Whether or not these laws and institutions actually have a positive impact on these outcomes can only be ascertained after they have been created or made sufficiently strong to work properly. In this context, evaluating the efficacy of the resources spent on such activities may not make much sense, since the impact will only be meaningful after this initial, necessary foundation-building stage. Supporting the writing of laws and the setting up of institutions such as inspectorates, citizens’ advocacy offices, and attorneys’ associations may simply be necessary investments, even if it is very difficult to know whether or not they have had, or will have, an impact on the ultimate outcomes that USAID wants to affect.

The one activity area within rule of law that might be amenable to randomized evaluation, at least in principle, is the support for rule of law-oriented nongovernmental organizations (NGOs). The problem here is that the preferred method of selecting NGOs for support is through a small grants competition, whereas a truly rigorous evaluation of the impact of support would require randomly choosing NGOs for funding. One possible solution would be to hold a small grants competition and, having ranked the applications from best to worst, work down the list funding every other one. Then, data would need to be collected on the quality of the performance and/or the impact in its area of focus of every NGO on the list—both those that were funded and those that were not—and a comparison could then be made across those groups. The problem, again, however, is to figure out what, precisely, to measure (which will

depend, in any case, on the particular goals that the NGO sets for itself). Also, unless the small-grants competition generates a very large number of high-quality applications, this method is not likely to generate very useful results. The need for a large number of funded and nonfunded NGOs will be increased by the likelihood that NGOs will propose different sets of activities, so “success” will have two possible sources—the difficulty of the tasks that the NGO sets out to accomplish and the benefits of having received the small grant—and the sample of NGOs analyzed will need to be large enough to permit the impact of funding through the “noise” of the random variation in task difficulty.

Selected Designs from Peru: Decentralization, Rule of Law, and Political Parties

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 subnational governments (such as “participatory budgeting”);
- strengthen the management skills of subnational governments in selected regions of Peru; and
- increase the capacity of nongovernmental organizations in these same regions to interact with their local government.

With the exception of some activities relating to national-level policies, all interventions under the program took place in seven selected subnational regions (also called departments): Ayacucho, Cusco, Huanuco, Junin, Pasco, San Martin, and Ucayali.⁴

These seven regions contain 61 provinces, which in turn contain 536 districts.⁵

Workshops on participatory budgeting, training of civil-society orga-

⁴As discussed elsewhere, the regions were nonrandomly 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 1980s.

⁵Peru 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.

nizations, and other interventions took place at the regional, provincial, and district levels.⁶

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.” This outcome is potentially measurable on different units of 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). Experimental designs could be used to study the impact of the decentralization program, and the cost of appropriately designed experimental evaluations could in fact be far beneath the actual costs spent on monitoring and evaluation.

Best-possible designs. We discuss best-possible designs from the perspective of program evaluation. First, we discuss what an ideal *ex ante* design for the decentralization program might have been in 2002, when the program was begun. Second, we 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.

A “*tabula rasa*” design. We assume that the decentralization program will be implemented in the seven nonrandomly 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. 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.⁷ Another is that district mayors and other actors may more easily become aware of treatments in neighboring districts. For both of these reasons, it may be useful to randomize instead at the provincial

⁶Relevant subnational authorities include members of regional councils, provincial mayors, and mayors of districts.

⁷Some interventions also occurred at the regional level, particularly toward the end of the program, yet these interventions constitute a relatively minor part of the program.

level. Then, all districts in a province that were 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.⁸

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 makes any difference.

Bundling the interventions may provide the best chance to estimate a causal effect of treatment.

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 receive *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 noncompliers in a medical trial. In this context, estimating the “effect of treatment on the treated” may be of interest.

It may be worth choosing pilot districts at random as well. In the first

⁸Standard errors may need to be adjusted to account for the clustering of treated districts within provinces.

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. Comparing municipal capacity across incorporated and unincorporated municipalities at the end of the pilot period may not lead to useful results; 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, 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 at least two 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 organizations (CSOs)? 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 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.⁹

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 cost no more than \$10,000, a small amount compared to the overall program budget.

In addition, with appropriate design modifications, there might be substantial net savings. One possibility for cost savings would involve substantially limiting the volume of output/outcome indicators gathered by each of the local subcontractors. For example, measures could be *sampled* across local jurisdictions, rather than gathered quarterly on each of 536 municipalities. A related idea is that local subcontractors could be asked to 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.

Other Examples: Rule of Law, Political Parties, and Extractive Industries

Several of the programs planned under the new Peru strategic assessment might also be amenable to randomized designs. In this section, we briefly review possibilities for experimental designs afforded by programs related to the rule of law, political parties, and extractive industries.

Rule of law. Most of the interventions under the rule of law programs implemented 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 (nonexperimental) evaluation plan that was actually adopted, cases administered by judges

⁹For reasons discussed above, it may also be useful to conduct the randomization at the provincial rather than district level.

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 nonexperimental design represented a valuable evaluation plan: There was a comparison made across treated and untreated units on an outcome measure of interest. In this and similar examples, the data seemed to show a substantial effect of treatment.

However, judges were nonrandomly 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.¹⁰ In principle, it would have been possible to assign district court judges to 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.

Political parties. One idea under the new political parties program is to provide assistance to the major national-level parties in opening or strengthening local offices in selected municipalities. At this point, however, the parties themselves would choose where to open offices, so the design is nonexperimental.

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 strengthening and an increase is found. Is this due to the effect of local-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.¹¹ The point is

¹⁰While 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 did not; this could have required pre- and postswitch data on both groups of judges. While still nonexperimental, 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.

¹¹In 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).

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 deemed feasible in this instance, it is possible in principle, and it would provide a stronger basis for impact attribution than a non-experimental approach.

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 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 conduct 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 among 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 rollout 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, for example, there were fewer marches and demonstrations in communities in which USAID worked than in those in which it did not work.

Selected Designs from Uganda: Civil Society, Parliamentary Strengthening, and Anticorruption

Large and Small Grants to CSOs

In the proposed project for Strengthening Democratic Linkages in Uganda Program, USAID proposes to provide at least \$100,000 per year for grants to CSOs to enable them to monitor local governments and help improve representation and service delivery at the local level.¹² These grants are thought to have two main effects: (1) to develop a more robust civil society by increasing the capacity of the CSOs who are awarded the grants, and (2) to improve the performance of government service delivery by increasing civic input and oversight of government officials.

Across carefully matched subcounties, large grants, small grants, and no grants will be allocated randomly to local CSOs working on HIV/AIDS. The goal is to compare the effects of large grants to CSOs (treatment group) versus small grants to CSOs (partial control group) in order to determine the effects of increases in CSO funding. Providing small grants to the partial control group allows USAID to assess independently the effect of greater monetary resources, while controlling for the nonmonetary effects of receiving a USAID grant (such as public recognition, special accounting requirements, and outside monitoring). It also facilitates the collection of equivalent data from CSOs in both the treatment and partial control groups. Both the treatment group and the partial control group will also be compared to CSOs in matching sub-counties where no grants are awarded (full control group) to evaluate the total effect of awarding a grant.

Carefully matched groups of three subcounties will be purposively selected so that the subcounties within each group are similar along a number of dimensions that are measurable and likely to be associated with CSO capacity and government service delivery for HIV/AIDS programs. Selection criteria might include the type, size, budget, and experience of the HIV/AIDS-related CSOs already working in the subcounties, as well as the subcounties' size, urban population, wealth, voting pat-

¹²The Strengthening Multi-Party Democracy in Uganda program also provides for \$100,000 per year for grants to CSOs, although for a somewhat different purpose.

terns, background of key officials, location, ethnic composition, number and type of health facilities, and infection rates. The most important criteria to ensure comparability should be determined in consultations with experts. Grouped subcounties might be next to each other but immediate proximity is not necessary (or even desirable).¹³

In each subcounty, one CSO working in HIV/AIDS will be selected with the aim of finding similar CSOs across three subcounties in the group. One subcounty in each group will be randomly assigned to receive a large CSO grant to monitor HIV/AIDS services in the subcounty. Another subcounty in the group will be randomly selected to receive a small CSO grant for HIV/AIDS. The remaining sub-county in the group will act as the pure control and receive no grant. This will be repeated for at least 50 groups, and preferably more.¹⁴ It is important to ensure that: (1) the large grant provides a significant increase to the existing budget of the CSOs, and that the small grants do not and (2) that the CSOs spend their grants entirely on HIV/AIDS activities within the selected subcounty and that there is not contamination (sharing of resources or expertise) across subcounties. It would probably work best to select CSOs that work only in a single subcounty to prevent the supplementing or siphoning off of funds to the treatment sites due to the grant. CSOs in both treatment and partial control groups should receive equivalent technical assistance and training on how to use the grant money and how to monitor and improve service delivery. USAID interactions with the CSOs in the treatment group, and partial control group should be equivalent throughout.

Evaluation. The primary question for evaluation purposes is: What are the effects of monetary grants on the organizational capacity of CSOs and on the ability of CSOs to monitor and improve government service delivery? The best possible evaluation for this type of project would be a large *N* randomized controlled field experiment. Because a large *N* study would require sizeable grants to at least 50 CSOs and additional monitoring and measurement, the costs are greater than that which is currently envisioned for CSO grants within the Linkages program. However, this design offers substantial benefits over a small *N* experiment and is of general interest to USAID.

¹³Instead of grouping subcounties in sets of three, it might be more feasibly to use an alternative stratified sampling procedure whereby all the subcounties in the sample are stratified into types according to key factors and then subcounties within each stratum are randomly assigned into each of the three categories.

¹⁴Depending on the districts chosen for Linkages, it may be possible to randomly select all the treatment and control subcounties from within the 10 districts.

Measurement. Data should be collected before the grants are awarded, after the money is given (or at several points during the grant period), and two years after the end of the grant in order to assess both short-term and medium-term effects of the monetary infusion. Equivalent data should be collected about CSOs and service delivery in the treatment, partial-control, and full-control subcounties. The ability of USAID to collect comparable data in the partial control group should be facilitated by the fact that the CSOs are receiving some funds from USAID. USAID may have to provide a small fee or incentive to the CSOs not receiving grants to enable the collection of similar intrusive and time-consuming data from the CSOs in the pure control group.

In order to study the effect of grants and increased resources on the organizational capacity of the CSOs, data should be collected on the budget, activities, operations, and planning of the CSOs. In addition, pre- and postintervention surveys can be conducted with CSO employees, volunteers, government officials and employees, and stakeholders to evaluate changes in the activities, effectiveness, and reputation of the CSOs.

In order to evaluate the effectiveness of grants' government service delivery data can be collected on HIV/AIDS services and outcomes within each subcounty. Much of these data may already be collected by the government (such as the periodic National Service Delivery Survey conducted by the Uganda Bureau of Statistics (UBOS)—though perhaps USAID would need to fund an oversampling in treatment and control subcounties) or perhaps it can be collected in collaboration with other donor projects such as the President's Emery Plan for AIDS Relief. Special attention should be given during the research design stage to determine the government activities that are likely to be affected by greater CSO involvement and how those activities might be accurately measured. Additional data collection could be done through surveys of service recipients or randomized checks on facilities and services. In addition, money-tracking studies of local government and government agencies could be conducted to evaluate the level of corruption in HIV/AIDS projects within the selected subcounties.

Possible alternatives

1. The grants could be given for an issue other than HIV/AIDS. Selected issues must be ones where (a) the government plays a major role in providing services and (b) there are measurable outcomes of service delivery.

2. The intervention can be carried out at either the district level or the village level instead of at the middle subcounty level. At higher levels of local government, CSOs are denser and better organized. While the ability

of CSOs to effect change in government may be greater at higher levels, the size of the grant needed to make a detectable difference will also be larger. Furthermore, it may be too difficult to find similar groups, and to protect units from contamination by other donors at higher levels of government.

3. If additional funds cannot be secured to conduct a large *N* randomized controlled experiment, a small *N* experiment could be conducted with the available funds, although with significantly less power to accurately evaluate the effects of CSO grants. In order to increase the number of possible comparisons, and to help control for the effect of context with a small number of treatment sites, a variation on the above design may be warranted. The inclusion of a second issue area may facilitate analysis in a small *N* context. For example, in each subcounty, one CSO working on education and one working on HIV/AIDS will be selected with the aim of finding similar CSOs across subcounty groups and issues. One subcounty will be randomly assigned to receive a large education grant and a small HIV/AIDS grant, and another subcounty will receive a large HIV/AIDS grant and a small education grant. Figure E-1 provides an illustration.

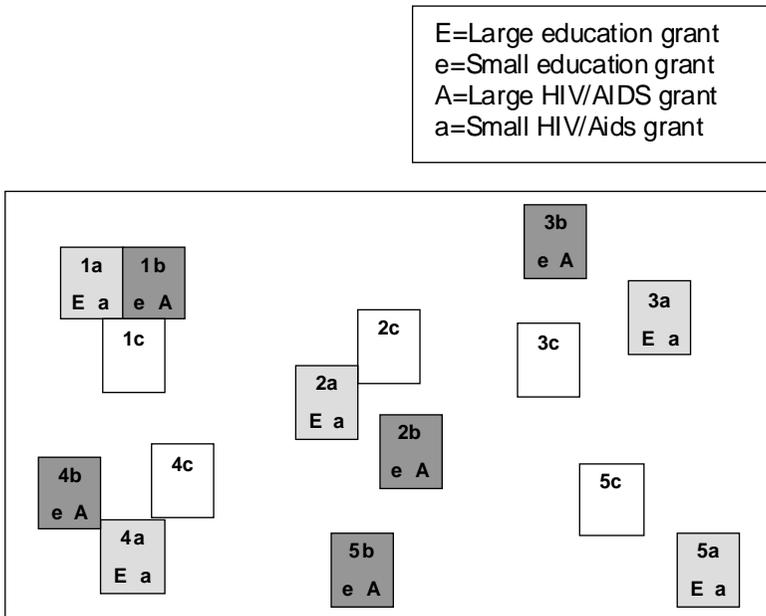


FIGURE E-1 Comparison of large and small grants to education and HIV/AIDS CSOs.

This research design affords several useful comparisons. Within a single subcounty, changes in the education CSO versus the HIV/AIDS CSO (one of which got a large grant and the other of which got a small grant) can be compared, and the degree of change in each sector can be evaluated. Within each subcounty group, the education CSOs (one with a large grant, one with a small grant, and one with no grant) can be compared and the changes in educational outcomes across the grouped subcounties can be compared. In addition, within each subcounty group, the two HIV/AIDS CSOs (one with a large grant, one with a small grant, and one with no grant) can be compared and the changes in HIV/AIDS outcomes across the grouped subcounties can be compared. The repetition of these comparisons across a number of different groups will help the researchers to parse out the effects of the grants from contextual factors.

Training and Assistance for a Random Selection of New Members of Parliament

The Strengthening Democratic Linkages in Uganda program seeks to enhance the knowledge, expertise, and resources of members of parliament (MPs) so they can more effectively operate in a multiparty parliament, legislate and perform oversight functions, foster sustainable development, and engage constituents, civil society, and local governments.

The entire group of new MPs (approximately 150) will be randomly divided into two groups. USAID can explain that they only have enough resources to work with half the group at a time and that the fairest way to decide is by lottery. To ensure that the partisan makeup of the treated group is equivalent to the control group, USAID will probably want to stratify by party affiliation. They may also want to stratify by other key factors such as previous political experience, committee assignment, and gender and randomly assign MPs within strata to ensure that the treatment and control groups are equivalent along critical dimensions.

The treatment group will receive intensive personalized training and assistance from technical personnel. This assistance may take the form of group trainings on key issues, weekly or bi-monthly individual meetings with trained legal assistants, regular research assistance on topics chosen by the MP, secretarial services, and/or repeated meetings with CSO representatives. The control group will not receive these additional services (at least initially). It is important to ensure that the intervention (1) is deemed useful by the MPs so that they continue to participate fully in the program for its duration; (2) is significant enough that the effects, if there are any, can be measured; and (3) is limited to the MPs in the treatment group alone and not easily passed on to those in the control group. For example, if the treatment was the distribution of a newsletter each week

to the treatment group, then it is very likely that many legislators in the control group would gain access to the newsletter and receive the same treatment as those in the treatment group.

Measurement. Jeremy Weinstein and Macartan Humphreys, in cooperation with the African Leadership Initiative, are currently producing annual scorecards for all of Uganda's MPs recording their behavior in the parliament, in committee, and in their constituencies. These scorecards could be used to compare the behavior of MPs in the treatment and control groups. In addition, surveys could be conducted with MPs to measure the knowledge and reported behavior of new MPs and to assess perceptions of fellow MPs. Surveys could also be conducted with parliamentary staff, civil service leaders, key stakeholders, or constituents to assess the reputation and influence of different legislators. Perhaps other measures of MP involvement (such as visits to the library) can be collected. Eventually, for those who run for reelection, the vote results could be used to evaluate popularity.

Evaluation. For the purposes of evaluation, the most important question is: What are the effects of technical training and assistance on the ability of individual legislators to operate more actively, effectively, and independently in parliament?

Possible alternatives

1. To reduce costs of the intervention, a smaller number of MPs can be selected to be in the treatment group. The required number depends on the intensity of the intervention, the quality of the measures, and the heterogeneity of the group, but a treatment group of 50 MPs may be sufficient.

2. If it is not politically feasible to provide benefits to only some of the new MPs, then the treatment could be conducted in a rollout fashion. Half (or one-third) of the MPs would receive the treatment for the first several years, and the other group would receive the treatment in the later part of the term. The interventions with each group would have to be timed to fit with the collection of data for the scorecards.

3. Returning MPs could also be included in the experiment, although returning MPs are more experienced and thus less likely to be affected by additional assistance. Their inclusion also adds to the heterogeneity of the population. The intervention activities (and the associated costs) would have to be greater, and/or more widespread, in order to discern an effect.

Revised Remuneration Policies to Fight Corruption

The Strengthening Capacity to Fight Corruption in Uganda Program suggests that “the Government of Uganda will consider increased pay for key personnel, through the implementation of an enhanced remuneration package for anti-corruption investigators and prosecutors.” The revised remuneration policies would “enable performance (job evaluation) based salary structures for anti-corruption prosecutors, investigators, and other officers within GOU entities such as the DEI, DPP and the CID fraud squad.”

The effects of changes in remuneration policies are of general interest to USAID. Although the implementation of the program cannot be manipulated to create contemporaneous control or comparison groups, the effects can still be evaluated effectively with a temporal comparison—before and after the intervention. The main consideration is to try to ensure that exogenous shocks do not take place during the period of measurement. For that reason we suggest that such an intervention could only be accurately evaluated if it took place some time before the other proposed reforms in the Request for Proposal for Strengthening Capacity to Fight Corruption in Uganda. Perhaps the changes in remuneration could be implemented immediately, while the other interventions are still in the planning stage.

Measurement. The main comparison is before the change in remuneration policies versus after the change. To evaluate the effect of changes in remuneration policies on recruitment and retention, the qualifications of the current employees will be assessed. In addition, the qualifications of all those who apply and former employees who sought alternative employment should also be assessed. To evaluate the effect of the remuneration policies changes on the effectiveness of anticorruption activities, the number of malpractices that are detected, effectively investigated, prosecuted, punished, and publicized before and after the changes can be compared.

Evaluation. The primary question from the perspective of evaluation is: How do changes in remuneration policies affect recruitment and retention of qualified personnel and the performance of employees?

Possible alternatives. If time permits, it would be better to stagger the changes in remuneration policies by types of civil servants or grades. For example, prosecutors could receive the new remuneration packages several months before the investigators. Thus, if there is an external shock, it is less likely to similarly affect the outcomes of every subject of the study.

**CURRENT AND RECENT USAID PROJECTS
AT THE TIME OF FIELD VISITS**

Albania (March 2007)

New

Local Government—RFP issued

Current/Recently Ended

Local Government (2004–end July 2007)—Urban Institute

Rule of Law (2004–end July 2007)—Casals

Political Parties and Civic Participation (2004–September 2007)—NDI/
IREX/Partners Albania

Anti-Corruption/MCC Threshold (2006-2008)—Chemonics

Peru (June 2007)

Current

Pro Decentralization (PRODES)—ARD, Inc.

Political parties/Elections—NDI/Transparencia

Congress Program—United Nations Development Program and
George Washington University

LAPOP Survey “Democracy Political Culture in Peru, 2006”—Vander-
bilt University

Not Included in Field Visit

Conflict Mitigation in Mining—CARE

Human Rights National Coordinator Institutional Development and
Therapy Attention to Victims of Torture and Political Violence—
Human Rights National Coordinator and Center for Psycho-Social
Attention

Trafficking in Persons—Capital Humano y Social Alternativo

Uganda (June 2007)

New

Democratic Linkages (within and among parliament, selected local gov-
ernments, and CSOs)—Center for Legislative Development SUNY
Albany

MCC Threshold (anti-corruption and civil society to improve procure-
ment systems and build capacity to more effectively investigate and
prosecute corruption cases)

Political parties and politically active CSOs (capacity building)—design-
ing project

Recent/Soon to End

Decentralization (to end December 2007)—ARD

Not Included in Field Visit

Community Resilience and Dialogue (September 2002–September 2007)—International Rescue Committee

CONSULTANT BIOGRAPHIES

Albania

Team Members: David Black, USAID; Rita Guenther, National Academies; Jo Husbands, National Academies; Karen Otto, consultant; Daniel Posner, consultant.

Karen Otto, a former USAID direct hire, is a monitoring and evaluation specialist/consultant with a strong background in democracy and governance (especially rule of law). She has developed 70 performance monitoring plans for proposals and ongoing development projects in a wide array of areas, particularly DG. She has evaluated the performance of many development projects and the operations of all federal courts in the United States, and has developed a formal evaluation system for the Administrative Office of the U.S. Courts to review courts under its jurisdiction. Ms. Otto has been a court administrator in federal, state, and municipal courts in the United States. She has been a rule of law advisor in USAID and a project manager for DG projects overseas. She has personal experience in many of the areas involved in DG activities: court administration (she was a court administrator), media (she was a journalist), judicial disciplinary system (she was an inspector in a judicial inspection service), etc.

Daniel Posner, associate professor of political science at the University of California, Los Angeles, conducts research in the following four broad areas: ethnic politics, ethnicity and economic development, political change in Africa, and social capital and civil society. His research in this area is motivated by a number of questions: When and why do some ethnic identities (and ethnic cleavages) matter for politics, and when do they not? Why, when people think about who they are, do they see themselves (and others) as members of particular ethnic groups, and why do the groups that they see themselves as part of have the sizes and physical locations that they do? How can we reconcile what we know about the fluidity and context dependence of ethnic identities and ethnic cleavages with the need to measure social diversity and code individuals by their

group affiliations? Why does ethnicity matter for collective action? How well are people able to identify the ethnic backgrounds of others? He approaches each of these questions with a combination of theory and the collection of original data (including experimental data).

Peru

Team Members: Moises Arce, consultant; Tabitha Benney, National Academies; David Black, USAID; Thad Dunning, consultant; Rita Guenther, National Academies.

Moises Arce is an associate professor in the Department of Political Science at the University of Missouri. His research focuses on the politics of market reform, comparative political economy, and Latin American politics (Peru). He received funding from the National Science Foundation, the Social Science Research Council, and the Fulbright Scholar Program. His publications include the book *Market Reform in Society: Post-Crisis Politics and Economic Change in Authoritarian Peru*, and articles in the *Journal of Politics*, *Comparative Politics*, *Comparative Political Studies*, and the *Latin American Research Review*. He previously taught at Louisiana State University. He received his Ph.D. in 2000 from the University of New Mexico.

Thad Dunning is assistant professor of political science and a research fellow at the Whitney and Betty MacMillan Center for International and Area Studies at Yale. His current research focuses on the influence of natural resource wealth on political regimes; other recent articles investigate the influence of foreign aid on democratization and the role of information technology in economic development. He conducts field research in Latin America and has also written on a range of methodological topics, including econometric corrections for selection effects and the use of natural experiments in the social sciences. Dunning's previous work has appeared in *International Organization*, the *Journal of Conflict Resolution*, *Studies in Comparative International Development*, *Geopolitics* and in a forthcoming *Handbook of Methodology* (Sage Publications). In 2006-2007, he was teaching an undergraduate lecture course and a seminar on ethnic politics and a graduate seminar on formal models of comparative politics. He received a Ph.D. in political science and an M.A. in economics from the University of California, Berkeley.

Uganda

Team Members: Mark Billera, USAID; Mame-Fatou Diagne, consultant; John Gerring, committee member; Jo Husbands, National Academies; Devra Cohen Moelher, consultant.

Mame-Fatou Diagne is a Ph.D. candidate in economics at the University of California, Berkeley. A native of Senegal, she graduated from the Institut d'Etudes Politiques de Paris and received a Master of International Affairs from Columbia University. She has worked as an emerging markets economist for Societe Generale in Paris and for Standard and Poor's in London, where she was the principal analyst for South Africa and other African-rated sovereigns. Her current areas of research are development, public and labor economics, and particularly, the economics of education and political economy in Africa.

Devra Cohen Moehler is an assistant professor of political science at Cornell University. She recently returned to Cornell from two years as a Harvard Academy Scholar at the Harvard Academy for International and Area Studies. Her research interests include political communications, education and democratization, consequences of political participation, political behavior, comparative constitution-making, law and development, cross-national survey research, and the international refugee regime. Her dissertation, based on research conducted in Uganda, focused on the effects of citizen participation in Ugandan constitution making in creating "distrusting democrats." She received her Ph.D. in political science from the University of Michigan and a B.A. in development studies from the University of California, Berkeley.

F

Voices from the Field: Model Questionnaire

Introductory Dialogue:

Good day. As you know, my name is _____. As part of ongoing attempts on the part of DCHA [the Bureau of Democracy, Conflict, and Humanitarian Assistance] to better understand the effect of our democracy promotion activities in countries around the world, we are conducting a series of surveys with DG advisors and activity managers. You have been selected to participate in this survey because of your extensive knowledge and experience. We will spend approximately 90 minutes with you asking a series of questions about your experiences. I will take handwritten notes of your responses. Please feel free to ask me clarifying questions as we progress. At the end of the interview, there will be an opportunity for you to address any subjects or issues that we may have missed or given less emphasis than they deserve. Please be assured that you can talk with candor; your responses will remain anonymous. We do intend to aggregate the responses of all our interlocutors for the purposes of reporting and improving DCHA recommended approaches in the future and we may use quotes from our interviews, stripped of identifying information. However, any specific references to what you tell us will only be used with your consent.

Do you have any questions before we begin?

Let's begin by talking about your work with USAID

1. In total, how long have you worked for USAID?
2. How much of your time with USAID has been spent working in the Democracy and Governance Sector?
3. In how many countries and for how long have you worked in the DG Sector with USAID? Please list for me the name of the country and how long you have worked in each country.
4. Which DG subsectors have you worked in for USAID? Please list the name of the country and the subsector(s) in which you worked in that country. *[Interviewer: Write the name of the country and place an X in the box below the subsector(s) for that country. (Subsectors: Civil Society, Rule of Law, Legislative Strengthening, Electoral Processes, Anti-Corruption, Media, Human Rights, Other)]*

Now let's talk about some of the specific USAID DG programs that you have worked on. First, we are interested in how you think about program success.

1. Considering all of the DG programs that you have worked on, supervised, or directly observed, can you tell me which one or two you think were the most successful?
2. *[Interviewer: If one program was identified above, skip this question and go directly to question, #7. If two programs are identified above, ask:]* In your view, which of these two was the most successful?
3. Let's get a little more information about this program. In which country was it carried out?
4. During which years did the program operate?
5. During this time, when were you involved with the program?
6. What was the approximate funding level? Please indicate the life of project funding and the annual funding.
7. What were the objectives of the program?
8. Can you please describe the basic operation of the program? How did it work?
9. And why do you say that this program was the most successful? What did it accomplish?
10. Can you give me a few examples of success?
11. Can you identify the particular factors that seem to have led to the success of this program and why each factor that you identify was important? *[Interviewer: Be sure to prompt informant to answer why each factor is important.]*
 Factor 1 & why important?
 Factor 2 & why important?

Factor 3 & why important?

Factor 4 & why important?

Factor 5 & why important?

12. Considering all of the factors that you have just told me about, can you identify which one or two are the most important contributors to the program's success?
 Most important factor
 2nd most important factor

We have developed a list of factors that have often been associated with program success and failure. Some of them are mirrored in the factors you have identified; a few others have not yet been mentioned. We would like you to describe for us how, if at all, these particular variables seem to be related to the success of the program.

13. Sometimes, program success can be influenced by country-specific enabling factors, things like the general level of economic development, cultural and social conditions, or historic precedent. On a scale of 1 to 5 with 5 representing the highest level of importance, how would you rank the importance of these factors in determining the success of the program? *[Interviewer: Circle a single number]*
14. Were any attributes in this cluster of factors particularly important, and if so, why?
 Attribute 1 & why important?
 Attribute 2 & why important?
15. Now let's look at the country more specifically in terms of democratic development. Sometimes political factors like level of commitment to reform, institutional capacity, level of corruption, level of press freedom, degree of political competition, capacity and activity of civil society, and other factors can influence the success of DG programs. On a scale of 1 to 5 with 5 representing the highest level of importance, how would you rank the importance of these factors in determining the success of the program? *[Interviewer: Circle a single number]*
16. Were any attributes in this cluster of factors particularly important, and if so, why?
 Attribute 1 & why important?
 Attribute 2 & why important?
17. Foreign policy priorities of the USG can sometimes have an important influence on program success. U.S. priorities in the country, the role of the Embassy, and other USG actors (DEA, DOD, CDC, MCC etc.) can affect the success of DG programs. On a scale of 1 to 5 with 5 representing the highest level of importance, how would you rank

the importance of these factors in determining the success of the program? *[Interviewer: Circle a single number]*

18. Were any attributes in this cluster of factors particularly important, and if so, why?
Attribute 1 & why important?
Attribute 2 & why important?
19. International factors often play a role in determining program success. The political conditions in the region, international political orientation and diplomatic considerations of the country, and the interests and activities of other donors might play varying roles. On a scale of 1 to 5 with 5 representing the highest level of importance, how would you rank the importance of these factors in determining the success of the program? *[Interviewer: Circle a single number]*
20. Were any attributes in this cluster of factors particularly important, and if so why?
Attribute 1 & why important?
Attribute 2 & why important?
21. Program-specific factors are also often important in determining success. Things like levels of funding for the program, length or sequencing of the program, implementation mechanism, quality of project design, quality or experience of the implementing partners' (contractors/grantees) staff or home office support; quality of the implementing partners' program management; quality of host country partners, willingness to take risks, etc., can all influence success. On a scale of 1 to 5 with 5 representing the highest level of importance, how would you rank the importance of these factors in determining the success of the program? *[Interviewer: Circle a single number]*
22. Were any attributes in this cluster of factors particularly important, and if so, why?
Attribute 1 & why important?
Attribute 2 & why important?
23. The USAID mission itself is often a factor associated with program success. For example, the priority given the DG sector, experience and staffing level of DG staff, programmatic relationships between DG and other mission sectors, the quality of mission management and leadership, and the impact of previous USAID activities can all be important. On a scale of 1 to 5 with 5 representing the highest level of importance, how would you rank the importance of these factors in determining the success of the program? *[Interviewer: Circle a single number]*
Attribute 1 & why important?
Attribute 2 & why important?

24. Now, let's look back briefly at the question where you identified a number of factors that you thought were determinants of the program's success. You mentioned *[Interviewer: Turn back to Question 15 and read a summary of each of the factors identified by respondent]*. Considering the factors that you mentioned and the factors that we have just discussed, would you like to make any additions or changes in the level of importance? Recall that the factors we have just discussed are: (1) Country-specific enabling environment, (2) Democratic/political, (3) Foreign policy/other donors, (4) International, (5) Program-specific, (6) USAID mission.

Most important factor

2nd most important factor

3rd most important factor

4th most important factor

5th most important factor

[Interviewer: If a second program was identified as successful, repeat the sequence of questions. If only one program was identified, go directly to the next series of questions.]

25. We have talked quite a bit about successful DG programs and it is nice to find out what works. Let's take a few minutes to consider the other side of the coin. Can you tell me about one or two of the biggest "turkeys"? As you reflect on your experience, what is the worst program that you ever worked with?
26. During which years did this program operate?
27. During this time, when were you involved with the program?
28. What was the approximate funding level? Please indicate the life of project funding and the annual funding.
29. What were the objectives of the program?
30. Can you please describe the basic operation of the program? How did it work?
31. And why do you say that this program not successful? Why was it a "turkey"?
32. If we think about the universe of factors we have discussed, can you identify which if any of the following factors contributed to the poor outcomes in this case and why? Recall that the factors we have discussed are: (1) Country-specific enabling environment, (2) Democratic/political, (3) Foreign policy/other donors, (4) International, (5) Program-specific, (6) USAID mission
- Factor 1 & why important?
- Factor 2 & why important?

Factor 3 & why important?

Factor 4 & why important?

Factor 5 & why important?

We are nearly done here! Thinking about program success and failure, everything else being equal, are there any type of programs or DG activities (rule of law, civil society, elections, parties and legislatures, anticorruption, decentralization, etc.) that you think are more likely or less likely to succeed than others? If so, which ones and why?

Sector 1 & why more or less likely successful?

Sector 2 & why more or less likely successful?

We would like to ask you to give us a few general observations and recommendations on the basis of your overall experience. First, what guidance would you give to a DG officer thinking about issues related to program sequencing and an appropriate or rational mix of programs in a DG portfolio?

1. Do you have any observations about the general characteristics of successful DG programs?
2. Do you have any additional comments or final observations?

THANK YOU VERY MUCH.
YOUR ANSWERS HAVE BEEN VERY HELPFUL.